END-TO-END ADMISSION CONTROL OF MULTICLASS TRAFFIC IN WCDMA MOBILE NETWORK AND WIRELINE DIFFERENTIATED SERVICES

XIAO LEI

NATIONAL UNIVERSITY OF SINGAPORE

2003

END-TO-END ADMISSION CONTROL OF MULTICLASS TRAFFIC IN WCDMA MOBILE NETWORK AND WIRELINE DIFFERENTIATED SERVICES

XIAO LEI

(B.Sc., Fudan University)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2003

Acknowledgements

I would like to take this opportunity to express my gratitude to all the people who have contributed to this thesis. Foremost among them are my supervisors, Dr. Wong Tung Chong and Dr. Chew Yong Huat. Both of them have given great guidance and advice in my study and research. I have learned enormously from them about the research, and as well as how to communicate with others.

I would also like to thank the other people in this project team, Nie Chun, Yao Jianxin and Govindan Saravanan, for the helps, discussions and suggestions to my research work.

Finally, I would like to thank my parents for their great love, encouragement and support in my two years studies.

Contents

Summary

In this thesis, we investigate the Quality of Service (QoS) provisioning issues of multiclass traffic across a Wideband Code Division Multiple Access (WCDMA) mobile network and a wireline Differentiated Services (DiffServ) Internet Protocol (IP) network, and focus on end-to-end admission control. The main objective is to propose an effective admission control algorithm for the end-to-end delivery of multimedia information between the mobile users and the fixed network users with specified QoS guarantees.

We define the mapping of QoS classes between the Universal Mobile Telecommunications Services (UMTS) and DiffServ networks according to different QoS requirements due to the different QoS architectures in the two domains. We propose a resource allocation and admission control scheme in DiffServ network and it inter-works with the four QoS classes in UMTS. Through the management of equivalent bandwidth allocation, the packet loss ratio at each router can be bounded; with a delay bound estimation, the statistical delay control at the router can also be obtained. Thus end-to-end packet loss ratio and statistical delay guarantee can be achieved. We also study the effect of buffer size on the four traffic classes with different priorities in the system.

We observe that the higher priority the traffic, the smaller the buffer size needed to provide the loss ratio guarantee. Only when the system is close to full utilization, the buffer size needed for a lower priority class, especially the background traffic with the lowest priority, increases drastically. The increase of buffer size of the real-time traffic imposes a small increase on packet queuing delay due to its high priority in the system. However, it can reduce the packet loss ratio significantly.

Furthermore, we apply the above scheme in end-to-end admission control, in both the UMTS DiffServ capable wireline network and the external DiffServ IP network. The admission region of the WCDMA cell is based on the outage probability and the packet loss ratio of each class in the wireless channel. Three different wireless admission region schemes, single-connection without retransmission, single-connection with retransmissions and multi-connection with retransmissions, are investigated. The wireless and wireline networks interact with each other in the end-to-end admission control. Only if both domains have enough resources to support the new and existing connections and the end-to-end QoS requirements can be guaranteed, then the new connection is admitted. Simulation results show that the schemes are effective in providing end-to-end QoS guarantees.

List of Figures

List of Tables

Glossary of Symbols

B	Buffer size
B_j	Buffer size of j^{th} class traffic
H	Hurst parameter of long range dependence source
$I_{\text{intercell}}$	Inter-cell interference
K_j	User Number of j^{th} class traffic
Q	Queue length in buffer
W_i	Waiting time of j^{th} class packet
$\overline{W_i}$	Mean waiting time of j^{th} class packet
b	Mean burst size
eb	Equivalent bandwidth
eb_j	Equivalent bandwidth of j^{th} class traffic
eb_j^k	Equivalent bandwidth of type j sources seen by type k traffic
m	Mean rate of aggregate sources
r	Transmission rate during On state of the source
γ	Asymptotic constant
γ_j	Asymptotic constant of j^{th} class traffic
δ	Asymptotic decay rate
ε	Packet loss rate
α	Transition rate from On to Off state of exponential on-off source
β	Transition rate from Off to On state of exponential on-off source

σ	Standard deviation
ρ	System utilization
ρ_i	Utilization of j^{th} class traffic
ω	Shape parameter of Pareto distribution
θ	Location parameter of Pareto distribution
μ	Mean value of exponential distribution

Abbreviations

AF	Assured Forwarding
ARQ	Automatic Repeat Request
ATM	Asynchronous Transfer Mode
BB	Bandwidth Broker
BCH	Broadcast Channel
BER	Bit Error Ratio
CAC	Connection Admission Control
CDE	Chernoff Dominant Eigenvalue
CDMA	Code Division Multiple Access
CN	Core Network
CPCH	Uplink Common Packet Channel
CS	Circuit-Switched
DCH	Dedicated Transport Channel
DiffServ	Differentiated Services
DPCCH	Dedicated Physical Control Channel
DPDCH	Dedicated Physical Data Channel
DS-CDMA	Direct-Sequence Code Division Multiple Access
DSCH	Downlink Shared Channel
DSCP	Differentiated Service Code Point
EB	Equivalent Bandwidth
EF	Expedited Forwarding

FACH	Forward Access Channel
FBM	Fractional Brownian Motion
FDD	Frequency Division Duplex
FEC	Forward Error Correction
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GSN	GPRS Support Node
GTP	GPRS Tunneling Protocol
GTP-U	GTP for the User Plane
HLR	Home Location Register
IETF	Internet Engineering Task Force
IMS	IP Multimedia Core Network Subsystem
IntServ	Integrated Services
IP	Internet Protocol
ISDN	Integrated Services Digital Network
MAI	Multiple Access Interference
ME	Mobile Equipment
MSC	Mobile Switching Center
MT	Mobile Termination
MTU	Maximum Transfer Unit
OVSF	Orthogonal Variable Spreading Factor
PCH	Paging Channel
PDF	Policy Decision Function
PDP	Packet Data Protocol

PDU	Protocol Data Unit
PHB	Per-Hop Behavior
PLMN	Public Land Mobile Network
PS	Packet-Switched
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RACH	Random Access Channel
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RRC	Radio Resource Control
RRM	Radio Resource Management
RSVP	Resource Reservation Setup Protocol
SBLP	Service Based Local Policy
SGSN	Serving GPRS Support Node
SIR	Signal to Interference Ratio
SNMP	Simple Network Management Protocol
SRNC	Serving RNC
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TE	Terminal Equipment
TOS	Type of Service
UDP	User Datagram Protocol
UE	User Equipment
UMTS	Universal Mobile Telecommunications Services
UTRAN	UMTS Terrestrial Radio Access Network

VBR	Variable Bit Rate
VLR	Visitor Location Register
WCDMA	Wideband Code Division Multiple Access

Chapter 1

Introduction

Wireless personal communications and Internet are the fastest growing segments of the telecommunication industry. Demand for high-speed wireless data and video services is expected to overtake voice services as the wireless industry grows and a hybrid wireless wideband CDMA/wireline Internet Protocol (IP)-based network will be the main platform for providing multimedia services to both mobile and fixed users.

As the end-to-end connection spans over both wireless wideband Code Division Multiple Access (CDMA) segment in the third generation wireless system and wireline IP-based network such as the Internet, the end-to-end Quality of Service (QoS) architecture consists of two parts: the wireless and the wireline QoSs. This proposed research will investigate how to interconnect a future wireless network with the IP network for seamless end-to-end information delivery.

1.1 QoS in Wireline Networks

Quality of Service is a broad term used to describe the overall performance experience that a user or application will receive over a wireless or wireline network.

QoS involves a broad range of technologies, architectures and protocols. Network operators achieve end-to-end QoS by ensuring that network elements apply consistent treatment to traffic flows as they traverse the network.

Despite the fast growth, most traffic on the Internet is still “best effort”, which means that all packets are given the same treatment without any guarantee in regards to the QoS parameters, such as loss ratio, delay and so on. However, with the increasing use of the Internet for real-time services (voice, video, etc.) and non-real-time services (data), there is a need for the Internet to provide different types of services having different QoS requirements.

There has been much research work done in the recent years on QoS issues in the Internet. The main QoS frameworks of interests include the Integrated Services (IntServ) [1] with Resource Reservation Setup Protocol (RSVP) [2] and the Differentiated Services (DiffServ) [3] which are defined by the Internet Engineering Task Force (IETF).

1.1.1 Integrated Services

The main idea of IntServ is to provide an application with the ability to choose its required QoS from a range of controlled options provided by the network. Its framework developed by IETF is to provide individualized QoS guarantees to individual application sessions. Thus it depends on the routers of the network having the ability to control the QoS and the means of signaling the requirements. RSVP provides the needed signaling protocol.

For the network to deliver a quantitatively specified QoS to a particular flow, it is usually necessary to set aside certain resources (e.g., bandwidth) for that flow. RSVP is an unidirectional control protocol that enables the QoS to be signaled and controlled. It helps to create and maintain resource reservations on each link along the transport path of the flow. With RSVP, the application call can signal the IP network to request the QoS level that it needs to provide the desired performance. If the network cannot provide the requested QoS level, the application may try a different QoS level, send the traffic as best-effort or reject the call.

Even though it can provide good QoS support, IntServ has the problem of scalability. This is because IntServ routers handle signaling and state management at the flow level to provide the desired QoS. If it is implemented at the Internet core network, it will place a huge burden on the core routers. In a very large network there are likely to be many users in the routers with similar QoS requirements, so it is much more efficient to use a collective approach to handle the traffic. This is performed by the differentiated services protocol to be described in the following paragraph.

1.1.2 Differentiated Services

Differentiated Services is defined by the IETF DiffServ Working Group to provide “relatively simple and coarse methods of providing differentiated classes of service for Internet traffic, to support various types of applications”. The main goal is to overcome the well-known limitations of Integrated Services and RSVP, namely, low scalability of per-flow management in the core routers.

DiffServ distinguishes between the edge and core routers. While edge routers process packets on the basis of a finer traffic granularity (e.g., per-flow), core routers only distinguish among a very limited number of traffic classes. Packets belonging to a given traffic class are identified by the bits in the DS field (a dedicated field in the header of IPv4 and IPv6 datagrams) of the IP packet header, and served by the routers according to a predefined per-hop behavior (PHB). In this way, traffic flows can be aggregated into a relative small number of PHBs which can be easily handled by core routers without scalability restrictions.

A PHB is a description of the rules a DiffServ compliant router uses to treat a packet belonging to a particular aggregated flow. Currently, some PHBs have been defined by the IETF, which include Assured Forwarding (AF) PHB [4] and Expedited Forwarding (EF) PHB [5].

1.2 QoS in Wireless Networks

Wireless networks provide communications to mobile users through the use of radio technologies, which include cellular system, cordless telephone, satellite communication, wireless local area network, etc. Compared with wireline network, the main advantages of wireless communications systems include: low wiring cost, rapid deployment of network and terminal mobility.

Current cellular communication systems are mainly used for voice-oriented services. Analog cellular systems are commonly referred to as the first generation systems. The digital systems currently in use, such as GSM, CdmaOne (IS-95) and US-TDMA, are second-generation systems. As for the third generation cellular systems, they are designed for multimedia communications, which means personal communications include not only voice, but also integrated services like image, video and data transmission.

The QoS framework for different service requirements should also be implemented in order to provide multimedia service capability in wireless networks.

Unfortunately, the quality of service over wireless transmission is much worse than that in the wireline networks. This is because the wireless medium has a much higher bit error ratio as a result of time-varying channel impairment. Wireless communication has the channel impairments such as multipath fading, background noise and multiuser interference. Thus it is more error-prone than that in the wireline networks. One of the solutions to this problem is to use Forward Error Correction (FEC), but it will add to the packet overhead and reduce the wireless channel capacity. Another alternative approach is to use Automatic Repeat Request (ARQ), which includes three main techniques, namely Go-Back-N, Stop and Wait and Selective Repeat. FEC is suitable for real-time traffic such as voice and video, while ARQ is suitable for non-real-time traffic like data. Go-Back-N ARQ is often used because of its simplicity and efficiency. Furthermore, a combination of both ARQ and FEC is possible. Because wireless medium is a shared medium, many users use the same channel for communication. The multiuser interference plays an important role in the channel performance. Since all the terminals use the wireless channel at the same frequency, a tight and fast power control is one of the most important aspects in the cellular CDMA systems.

While wireless communication provides the user with mobility support, it also brings about the problems of service performance degradation and call handover. In order to provide higher capacity, the wireless system needs to set up more microcells architectures. It results in more frequent handovers of mobile users and higher call blocking and dropping probability, which degrade the quality of service in the system.

1.3 Contributions of Thesis

This thesis presents and studies the admission control for a resource allocation scheme with multiclass traffic in a DiffServ IP network. The four QoS classes of WCDMA (UMTS) system is mapped and used as the traffic sources in the DiffServ network.

We further extend this scheme to end-to-end admission control from the third-generation wireless system (WCDMA) to the wireline (DiffServ) networks. This thesis also investigates how to interconnect a future wireless network with the Internet for seamless end-to-end information transmission. The simulation results show that this scheme can satisfy both the packet loss ratio and end-to-end delay requirements for the multiclass traffic.

1.4 Organization of Thesis

This chapter gives an overview of the QoS issues of the wireline and wireless communication networks. Chapter 2 describes the framework of a DiffServ network and gives a survey on the available admission control schemes. Chapter 3 introduces the WCDMA system, Universal Mobile Telecommunications Services (UMTS) and the wireless interface. In chapter 4, we present the DiffServ network admission scheme proposed and its simulation results. We investigate the end-to-end admission control scheme from the wireless network to the wireline network in chapter 5. Finally, the thesis is concluded in chapter 6.

Chapter 2

Differentiated Services Network

There is a need to provide different levels of QoS to different traffic flows as the amount of traffic traversing through the Internet grows and the variety of applications increases. The Differentiated Services framework is designed to provide a simple, easy to implement and low overhead frame structure to support a range of network services that are differentiated based on per-hop behaviors (PHBs).

2.1 Differentiated Services Architecture

In a DiffServ network, the routers classify a packet into one aggregate traffic type for processing. This is implemented by marking the Type of Service (TOS) in IPv4 header or Traffic Class (IPv6) as the Differentiated Service Field (DS Field) in an IP packet. This is an 8 bits field, but the first 6 bits become the DS code point (DSCP) field, and the last 2 bits are currently unused. The DSCP field of all packets will be checked at each DS-compliant router. The router will classify all the packets with the same field into a single class or behavior aggregate and select the appropriate PHB from a predefined table. Thus only a small number of aggregated flows are seen in the DiffServ network instead of numerous individual flows.

2.1.1 DiffServ Network Domain

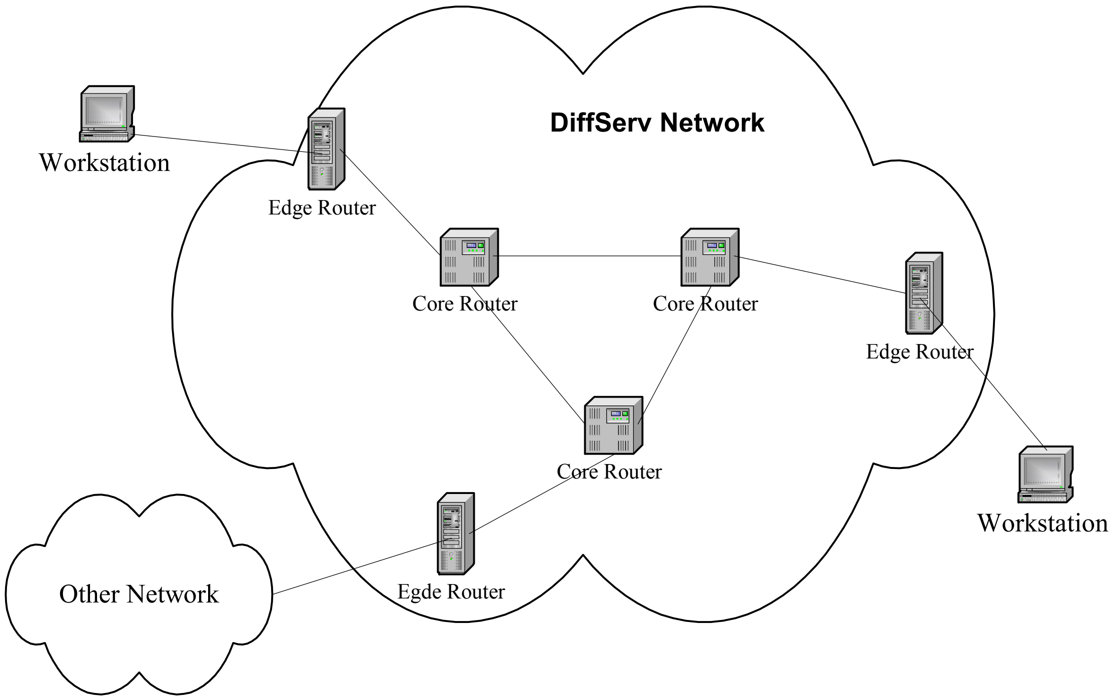


Figure 2.1: DiffServ Network Domain

A DiffServ network model is given in Figure 2.1. We divide the routers in the DiffServ network into two categories, namely Edge Routers and Core Routers, according to their characteristics and functions. Edge routers are at the boundary of the DiffServ domain and interconnect the domain to other adjacent networks or end users, while core routers only connect to other core routers or edge routers within the same DiffServ domain. Edge routers are responsible for classifying packets, setting DS bits in packets, and conditioning packets for all the incoming flows, while core routers efficiently forward large bundles of aggregate traffic at high speed.

When a packet comes to a DiffServ network, the classifier at the edge router systematically groups the packet based on the information of one or more packet header fields and the marker sets the DSCP field appropriately. This field identifies the class of traffic (behavior aggregate) the packet belongs to. The traffic conditioner performs the traffic conditioning functions such as metering, shaping, dropping and remarking. A DiffServ profile is a description of the traffic characteristics of a flow such as rate and burst size. In general, each packet is either in-profile or out-of-profile based on the metering result at the arrival time of the packet. In-profile packets obtain better traffic conditioning and forwarding treatment than out-of-profile packets. The shaper delays some or all of the packets in a traffic stream to change the traffic profile to a more conforming traffic characteristics, while the dropper discards some or all of the packets in a traffic stream to ensure that it conforms to the desired traffic profile. This process is known as policing the flow traffic.

The core routers will check the DSCP of every incoming packet inside the DiffServ network and determine its per-hop behavior. Since only edge routers store the individual flow information, the core routers do not need this information, the DiffServ networks have good performance in terms of scalability.

2.1.2 Per-Hop Behavior

A per-hop behavior (PHB) is a description of the forwarding behavior of a DS node applied to a particular DS behavior aggregate [3]. The PHB is the means by which a router allocates resources to behavior aggregates, and the differentiated services architecture is constructed under the basis of such hop-by-hop resource allocation mechanism. Currently, DiffServ working group has defined a number of PHBs and recommends a DSCP for each one of them. These include Expedited Forwarding PHB and Assured Forwarding PHB group.

Expedited Forwarding (EF) PHB can be used to build a low loss, low latency, low jitter, assured bandwidth, end-to-end service through DiffServ domains. It is generally described as the Premium service. The dominant causes of delay in packet networks are propagation delays on the links and queuing delays in the switches and routers. As the propagation delays are fixed property of the network topology, delay will be minimized if the queuing delays are minimized. In order to minimize the queuing delay for an EF packet, it should see small or no queues in the system when it comes to the routers or switches. Thus it is necessary to ensure that the service rate of EF packets at a router exceeds their arrival rate over long and short time intervals and is independent of the load of other (Non-EF) traffic. A variety of scheduling schemes can be used to realize the EF PHB, and a priority queue is considered as the canonical example of an implementation.

Assured Forwarding (AF) PHB group is a means to provide different levels of forwarding assurances for IP packets in a DiffServ domain. Four AF classes are defined, and each AF class in a DiffServ router is allocated a certain amount of forwarding resources such as buffer and bandwidth. Within each AF class, IP packets are marked with one of three drop precedence values as shown in Table 2.1. The DiffServ router will protect packets with a lower drop precedence value from being lost by preferably discarding packets with a higher drop precedence value when congestion occurs.

Table 2.1: Assured Forwarding PHB

	Class 1	Class 2	Class 3	Class 4
Low Drop Precedence	AF11	AF21	AF31	AF41
Medium Drop Precedence	AF12	AF22	AF32	AF42
High Drop Precedence	AF13	AF23	AF33	AF43

2.1.3 DiffServ Network Provisioning

The main objective of network provisioning is to enhance the performance of a network and improve its quality of service. Network provisioning consists of two parts: traffic management and resource allocation. Traffic management involves the regulation of the flow traffic through the network such as traffic conditioning at the edge router and congestion control in the network. Resource allocation deals with the resource management in the network which includes link bandwidth, buffer space, etc. In fact, traffic management and resource allocation are intertwined rather than independent from each other. An efficient and adaptive network provisioning scheme is one of the main challenges in the issues of network QoS.

DiffServ network provisioning is still under research. Currently, it is mainly realized by static provisioning of network resources. Jacobson and Nichols [6] proposed the concept of a Bandwidth Broker (BB) which is an administrative entity residing in each DiffServ domain. The BB has two responsibilities, one is the intra-domain resource management and the other is the inter-domain service agreement negotiation. The BB performs resource allocation through admission control in its own domain. On the other hand, the BB negotiates with its neighbour networks, sets up bilateral service level agreement and manages the adequate intra-domain resource allocation in providing end-to-end connection QoS. When an allocation is desired for an incoming flow, a request is sent to the BB. This request includes service type, target rate, burst size, and the service time period. The BB first authenticates the credentials of the requester, then checks whether there are sufficient unallocated resources to meet the request. If the request passes all the tests, the available resource is allocated to the requester and the flow specification is recorded.

2.2 Admission Control

Connection admission control (CAC) evaluates whether the network can provide the requested service to the coming new flow while maintaining the service promised to the other existing flows. From the QoS requirement and traffic characteristics of the incoming flow, resources demanded by the flow are determined. From the QoS and traffic characteristics of the admitted flows in the network, allocated resources are determined. If the remaining resources are not less than the requested resources needed by a new flow, the service can be provided and the flow will be admitted. If the request is rejected, renegotiations may be performed for a less stringent traffic profile or QoS requirement. The best effort service is the lowest priority class to be provided.

There has been much research work done in the area of admission control. Some schemes can provide deterministic guarantee (hard guarantee) QoSs, while others only provide statistical guarantee (soft guarantee) QoSs. The hard guarantee scheme is too conservative and the soft guarantee scheme is more flexible and can increase the network capacity. In general, many of the admission control policies can be classified as measurement-based CAC, resource allocation-based CAC and hybrid CAC.

2.2.1 Measurement-Based CAC

There is now an increasing interest in measurement-based admission control. Using measurement-based scheme, routers periodically collect measured results of necessary quantities representing the state of the network such as the available bandwidth on a link. Admission decisions are then based on these measurements rather than on worst-case bounds.

In [7,9], a probing packet stream with the same traffic parameters of a requesting connection is sent from the sender to the receiver hosts. The packet loss ratio, delay and other QoS metrics are measured at the receiver. These are used to describe the congestion level of the network. If the measurement result is acceptable, the connection is admitted. Otherwise it is rejected. A similar idea is considered by Bianchi and Blefari-Melazzi [8]. In their scheme, a packet is sent with a medium drop precedence for every AF class to get the congestion information of the lower drop precedence traffic. This can be inferred as the medium drop precedence packet will be dropped first if congestion of the lower drop precedence traffic is detected in the network. Since the final admission decision is made at endpoint nodes, these schemes are classified as Endpoint Admission Control.

Knightly and Qiu [10] employ adaptive and measurement-based maximal rate envelops of the aggregate traffic flow to provide a general and accurate traffic characterization. This characterization captures its temporal correlation and the available statistical multiplexing gain. Both the average and variance of these traffic envelops, as well as a target loss rate, are used as the input parameters for the admission algorithm. The authors also introduce the notion of schedulability confidence level to describe the uncertainty of the measurement-based prediction and reflect temporal variations in the measured envelop. In [11], Oottamakorn and Bushmitch present a CAC based on the measurement of global effective envelops of the arriving aggregate traffic and the service curves of their corresponding departing aggregate traffic.

While measurement-based CACs can achieve a higher network utilization, in general they can only provide statistical guarantees and are practical only for highly predictable traffic or large traffic aggregations. If deterministic QoS guarantees (e.g. delay and loss ratio) need to be achieved, a measurement-based admission scheme may fall short of the task.

2.2.2 Resource Allocation-Based CAC

For resource allocation-based admission control, the general description of the scheme is as follows. When a new connection request arrives, it sends the request message to the network, together with its traffic parameters and QoS (e.g., delay and loss ratio) requirements. The network will calculate the available resources, such as the bandwidth on each link. If there is a path available for the new connection and it can provide the necessary QoS, the request will be accepted.

In a DiffServ network, the Bandwidth Broker (BB) will be the admission control agent for the whole domain. The BB should have a database with information about the network topology, connections, links and routers status. This removes the need for core routers to store the individual connection information. The BB is responsible for all the admission control decisions and the network resource allocation. The available resource is calculated through the information stored in its database. For a large network, if the whole domain information storage and admission control are hard for one network element to handle, distributed mechanism can be used. QoS routing plays an important role in this architecture. Its function is to find a suitable (optimal) path from the source to the destination which can provide the required QoS. In general, optimal routing with multiple QoS metrics is an NP-complete problem, some heuristic methods such as Multiple Constraints Bellman-Ford [13] are available. They can provide both bandwidth and delay constraints.

In [12], Zhang and Mouftah introduce a sender-initiated resource reservation mechanism over DiffServ network to provide end-to-end QoS guarantees. This is similar to RSVP. Agrawal and Krishnamoorthy [14] present an algorithm for identification of critical resources in the differentiated service domain, and the resource provisioning on this domain is based on these critical resources under some given survivability constraints for robustness.

The most important problem of resource allocation-based CACs is how to calculate the occupied and available resources such as link bandwidth in the network. Equivalent Bandwidth (EB) has been widely researched in the Asynchronous Transfer Mode (ATM) and IP environment. It is one of the main approaches in this thesis. A detail introduction about EB will be presented in a later chapter.

2.2.3 Hybrid CAC

Resource Allocation-based CACs provide accurate QoS bounds at the expense of network utilization as well as increased processing load at the central admission control entity. In fact the routers can directly estimate the number of connection in a given class on a link from the measured load by dividing the load by the sustainable connection rate for that class. The larger the number of connections, the more accurate the estimation is. This is the main idea of hybrid admission control. This type of scheme can provide satisfying results if the number of connections is large enough to diminish the imprecision due to traffic fluctuations.

2.2.4 Summary

These three types of CAC schemes have quite different characteristics and are expected to give different results. Resource allocation-based CAC performs most conservatively and enforces the constraints at the expenses of network utilization. On the other hand, measurement-based CAC does not always satisfy the constraints, but the utilization (hence the number of accepted connections) is higher than that of the former. The utilization of hybrid scheme is slightly larger than that of resource allocation-based CAC because it estimates the allocation using measurements.

Of all the admission control schemes surveyed, we discover that very few of them deal with the problem of providing both packet loss and delay guarantees with simple algorithms. Furthermore, most of them consider little about the multiclass services environment such as in a DiffServ network.

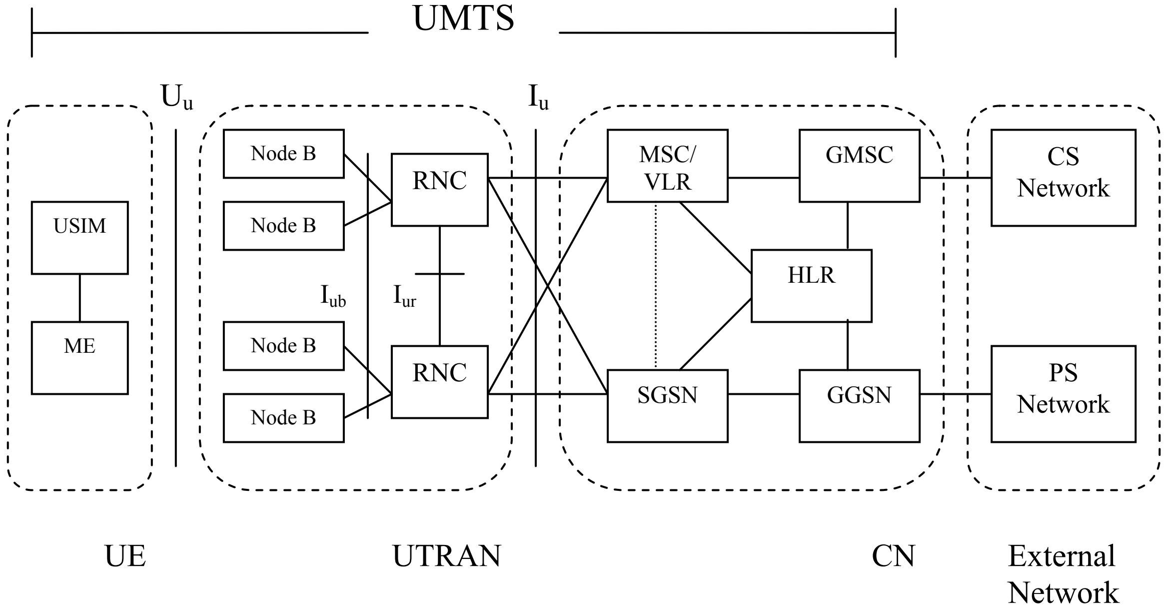
Chapter 3

WCDMA and UMTS

Universal mobile telecommunication system (UMTS) is a third-generation mobile communication system, designed to provide a wide range of applications, or more generally, to provide most of the services those are now available to fixed network customers to mobile users.

In an UMTS system, Wideband Code Division Multiple Access (WCDMA) [15] is the main air interface specified. WCDMA is a wideband Direct-Sequence Code Division Multiple Access (DS-CDMA) system. User information bits are spread to a wider bandwidth by multiplying the user data with high rate pseudo-random bits (called chips), which is also known as the CDMA spreading codes. WCDMA is anticipated to provide the third-generation mobile communication system with the high flexibility to support high rate (e.g., up to 2 Mbps) multimedia services. WCDMA uses Frequency Division Duplex (FDD) mode to support the uplink and downlink traffic. The multi-rate services are realized through the use of variable spreading factors and multi-code transmission. Both Circuit-Switched (CS) and Packet-Switched (PS) traffic are supported in WCDMA.

3.1 UMTS Architecture



U_u UMTS air interface
I_u Interface between RNC and CN
I_{ub} Interface between Node B and RNC
I_{ur} Interface between two RNCs

Figure 3.1: Network elements in a PLMN

Figure 3.1 shows the architecture of a complete Public Land Mobile Network (PLMN), in which the left portion is UMTS, and the right portion is the external network connected to UMTS. These external networks can be divided into two groups: Circuit-Switched networks (e.g., PSTN and ISDN) and Packet-Switched networks (e.g., Internet or DiffServ IP network).

The UMTS system consists of a number of logical network elements, each has a defined functionality. Functionally the network elements are grouped into: (1) User Equipment (UE) that interfaces the user and the radio interface, (2) UMTS Terrestrial Radio Access Network (UTRAN) which handles all radio-related functionality, and (3) Core Network (CN) that is responsible for switching and routing calls and data connections to external networks.

UE consists of two parts: Mobile Equipment (ME) which is the radio terminal used for radio communication over the U_u interface and UMTS Subscriber Identity Mobile (USIM), a smartcard holding the subscriber information.

UTRAN also consists of two distinct elements: one is Node B which converts and exchanges the data between the I_{ub} and U_u interfaces and participates in radio resource management, the other is Radio Network Controller (RNC) that controls the radio resources in its domain, and it is also the service access point for all services UTRAN provides to the CN.

The Core Network consists of two domains: Circuit-Switched domain and Packet-Switched domain. The Circuit-Switched domain centers around the Mobile Switching Center (MSC) and the Visitor Location Register (VLR). The Gateway MSC (GMSC) is the switch at the point where UMTS is connected to external CS networks; the Packet-Switched domain centers on the GSN (GPRS Support Node), and the Serving GPRS Support Node (SGSN) functionalities are similar to those of the MSC and VLR but are for Packet-Switched services. The Gateway GPRS Support Node (GGSN) functionality is similar to that of GMSC but it connects to external PS networks. The Home Location Register (HLR) is a database located in the user's home system that stores the information of user's service profile.

General Packet Radio Service (GPRS), developed as the packet-switched extension of the GSM network to enable high-speed access to IP-based services, is the foundation for the packet-switched domain of the UMTS core network. From the Release 5 of 3GPP specification, IP Multimedia Core Network Subsystem (IMS) [17] is introduced to support IP multimedia services.

The IMS comprises all core network elements for provision of multimedia services based on the session control capability defined by IETF and utilizes the PS domain. The IMS is designed to be conformant to IETF Internet standards to achieve access independence and maintain a smooth interoperation with wireline terminals across the Internet. This enables PLMN operators to offer multimedia services based on Internet applications, services and protocols for the wireless users such as voice, video, messaging, data and web-based technologies.

3.2 WCDMA Radio Interface

In WCDMA Frequency Division Duplex (FDD) mode, its uplink frequency band will be mainly deployed around 1950 MHz and downlink band is around 2150 MHz. The spacing between individual transmission channels is about 5 MHz, and the chip rate is 3.84 Mcps.

3.2.1 Spreading and Scrambling

Transmissions from one single channel are separated by the spreading (channelization) codes such as the dedicated physical channel in the uplink from one mobile station and the downlink connections from one base station. The spreading code used in UTRAN is Orthogonal Variable Spreading Factor (OVSF), which allows the spreading factor to be changed while the orthogonality between different codes of different lengths is maintained.

In addition to separating the different channels from one source, there is also the need to separate mobile stations or base stations from each other, and the scrambling code is used to implement this function. Scrambling is used on top of the spreading and it does not spread the bandwidth of signal. Thus it only makes the channels from different sources to be separated. In the uplink channels, short and long scrambling codes can be used, while in the downlink channels, only the long codes (Gold Code) are used.

3.2.2 Transport and Physical Channel

In UTRAN, the user data from high layers is transmitted in the transport channels over the air. These transport channels are mapped to the physical channels in the physical layer. The physical channel can support variable bit rate transport channels and multiplex several services to one connection.

There are two main types of transport channels: dedicated channels and common channels. The dedicated channel is reserved only for single user, while the common channel can be shared by multiple users in a cell. There is only one type of dedicated transport channel (DCH) which has the features such as fast data rate change, fast power control and soft handover. For common transport channel, six types are defined in UTRAN which includes Broadcast Channel (BCH), Forward Access Channel (FACH), Paging Channel (PCH), Random Access Channel (RACH), Uplink Common Packet Channel (CPCH) and Downlink Shared Channel (DSCH). Common channels do not have soft handover but some of them have fast power control. There are three types of transport channel that can be used for packet transmission in WCDMA: dedicated (DCH), common (RACH, FACH, CPCH) and shared (DSCH) transport channels.

The physical channels carry the information only relevant to the physical layer. The DCH is mapped to two physical channels, the Dedicated Physical Data Channel (DPDCH) carries the high layer information, i.e., the user data, while the Dedicated Physical Control Channel (DPCCH) carries the control information in the physical layer. The user data can be transmitted on a DPDCH with a possible spreading factor ranging from 4 to 256, which is shown in Table 3.1. If high data rate are needed, parallel code channels can be used, and the maximum number is six.

Table 3.1: Uplink DPDCH Data Rates

DPDCH spreading factor	DPDCH channel rate (kbps)	Maximum user data rate with ½ rate coding (kbps)
256	15	7.5
128	30	15
64	60	30
32	120	60
16	240	120
8	480	240
4	960	480

3.2.3 Power Control

Fast and efficient power control is one of the most important aspects in WCDMA, especially in the uplink direction when single user detectors are used. Because all the users in one cell use the same frequency to transmit information, a single overpowered user can block a whole cell due to the near-far problem of CDMA systems. The solution to this is to equalize the received power of all mobile stations by power control.

Though the open-loop power control is used in WCDMA, it can only provide a coarse initial power setting of the mobile station at the beginning of the connection. The fast closed-loop power control is applied in WCDMA after connection has been set up. For example, the base station performs frequent measurement of the received

Signal-to-Interference Ratio (SIR) of every user and compares it with a target value. If the measured SIR is higher than the target value, the base station will inform the mobile station to decrease its transmission power. If it is lower, the base station will inform the mobile station to increase its transmission power. Thus fast power control can remove the problem of imbalance of received powers at the base station.

3.3 UMTS Quality of Service

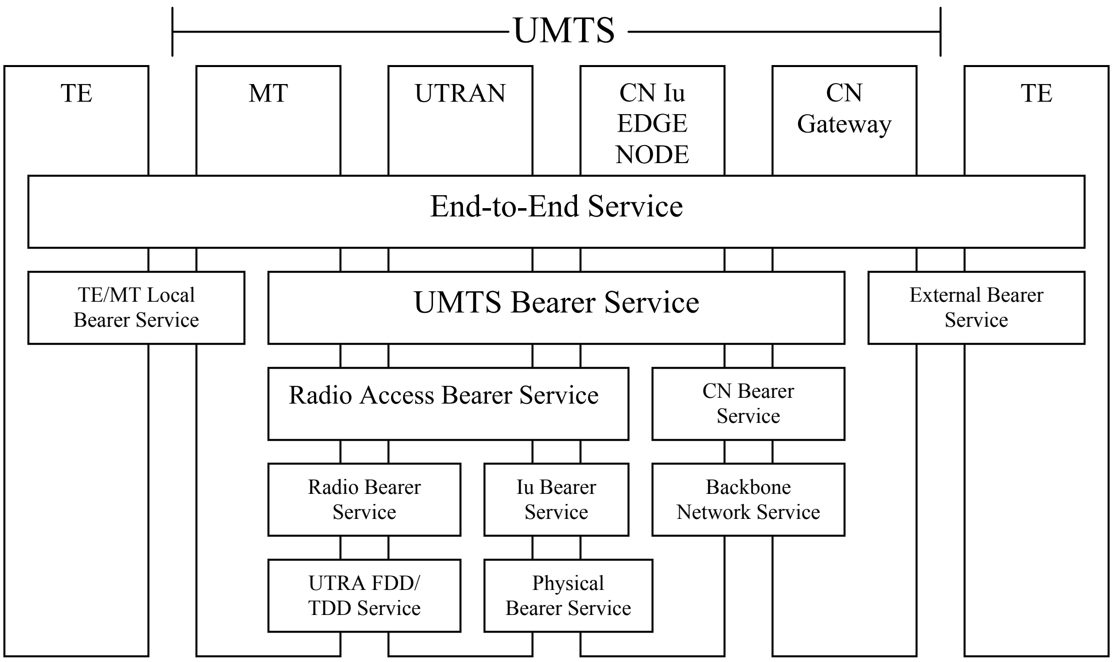


Figure 3.2: UMTS QoS Architecture

Figure 3.2 shows the UMTS QoS service architecture from [16] which is the same as with the QoS definition of GPRS Release 99. The end-to-end service QoS requirement is provided by: (1) the TE/MT Local Bearer Service between the Terminal Equipment (TE) and the Mobile Termination (MT), (2) the UMTS Bearer Service which includes the MT, UTRAN, CN Iu Edge Node and the CN Gateway, and (3) the External Bearer Service between the CN Gateway and the TE.

3.3.1 UMTS QoS Classes

There are four QoS classes defined by 3GPP in UMTS architecture:

- Conversational class;
- Streaming class;
- Interactive class;
- Background class.

The main difference between these QoS classes is how delay sensitive the traffic is. Conversational and Streaming classes are mainly intended to be used to carry real-time traffic. Real-time conversational services such as voice over IP, are the most delay sensitive applications and these data streams should be carried in Conversational class.

Interactive and Background classes are mainly to be used by Internet applications like WWW, Email, Telnet, and FTP. With looser delay requirements compared to conversational and streaming classes, both Interactive and Background classes provide better error rate by means of channel coding and retransmission. The main difference between Interactive and Background classes is that the Interactive class is used by interactive applications, e.g., Web browsing, while the Background class is meant for background traffic, e.g., background download of emails or files. Responsiveness of the interactive applications is provided by separating interactive and background classes. The Conversational class traffic has highest priority in the packet scheduling and the Background class traffic has the lowest priority. Thus background applications use transmission resources only when applications of the other three classes do not need them.

3.3.2 UMTS QoS Management

The QoS management functions of all the UMTS network elements together will provide the provision of negotiated services of the UMTS bearer service, and these functions are divided into two groups: management functions in the control plane and management functions in the user plane.

The control plane includes these functions:

- Service Manager: Coordinating the functions of control plane.
- Translation Function: Converting between the UMTS bearer service control and the service control of the interfacing external networks such as DiffServ IP network.
- Admission/Capability Control: Maintaining all the information of UMTS network resources. It verifies the available resource and decides whether to accept the request and allocate the corresponding resources.
- Subscription Control: Verifying the administrative rights for the use of a bearer service with specified QoS attributes.

The user plane includes these functions:

- Classification Function: Assigning each data unit received from external network or local bearer service to the appropriated UMTS bearer service according to its QoS requirement.
- Traffic Conditioner: Performing traffic policing or shaping to ensure that it conforms to the negotiated QoS.
- Mapping Function: Marking each data unit with the specified QoS indication of the bearer service.

- Resource Manager: Distributing the resources between all the bearer services according to QoS requirements by means of scheduling, bandwidth allocation, etc.

3.4 Admission Control in WCDMA

In a power-controlled CDMA system, there is no absolute limits on the number of users that can be supported in each cell. However, if the air interface load is not controlled, the users could increase excessively. Then the QoS of existing users will degenerate. The general capacity of a CDMA system is given in [18]. We can see that it is determined by many factors, e.g., the chip rate, the data transmission rate, the required SIR value and the inter-cell interference factor.

Since the WCDMA air interface adopts the FDD mode, the downlink and uplink transmissions can be considered independently and the traffic load in both links can also be asymmetric. The radio access bearer is admitted into the system only if both uplink and downlink admission control requirements are fulfilled. In the downlink direction the power level must consider two factors. One is the power emitted by the base station and the other is the limit to the powers used for each individual channel. In the uplink direction the emitted power levels on each channel need to be considered. The SIR is generated by the Multiple Access Interference (MAI) by all the uplink signals and the propagation conditions. So the MAI (or SIR) and power limitation are the basic criteria for the admission control algorithm in WCDMA.

The system load factor can be used as the criteria for admission control [19]. If the load factor becomes close to 1, the CDMA system has reached its capacity. This simple CAC algorithm is to derive an average cell capacity in terms of the number of connections so that a connection-based CAC can be adopted. In such a scheme, the load of a cell is formulated as the weighted sum of the number of active users in the reference cell and its neighbouring cells.

Another approach to the CAC problem is based on the idea that the CDMA capacity is strictly related to power limits which prevents the power control to reach a new equilibrium when the load is too high. In fact, the power levels are increased by the power control mechanism when the interference increases to keep the SIR at a target value. As a result, the level of power emitted with respect to the limit can be adopted as a load indicator in the admission decision. The decision rule can be quite simple like considering a threshold on the emitted power and admitting new calls only if the powers considered are below the threshold [20-22].

The total interference level at the base station can be adopted as a load measure in the admission procedure [23]. The base station will make periodic measurement of the SIR value of every existing user in their uplink transmissions. A threshold value is set by the admission control scheme when there is a connection request. The base station will check the SIR value of users in the cell. If the measured SIR is larger than the threshold, the call is accepted. Otherwise, it is rejected.

Several papers have presented schemes to implement the prioritized admission control algorithms [19,20,22]. The main idea is to let higher priority classes have privilege over lower priority classes at call admission, i.e., the more important calls have higher chance of being accepted. Higher priority classes can also preempt the lower classes, if the required QoS of higher classes are not satisfied. The lower priority traffic classes will release their bandwidth to higher priority class by possibly sacrificing their delay or loss constraints when necessary.

However, the above admission control schemes do not consider the problem of outage probability which has great effect on the connection level QoS. All of them only consider the admission problem in WCDMA air interface. Since the end-to-end connection spans over both the wireless and wireline networks, how to provide the seamless connection between them is still an issue to be investigated.

Chapter 4

DiffServ Network Admission Control

The objective of IETF DiffServ Working Group is to define a simple framework and architecture to implement DiffServ services. The additional network management and provisioning mechanisms such as admission control are under further research and implementation by the operators.

In this chapter, we propose an admission control scheme in the DiffServ network which attempts to work with the four QoS classes defined in UMTS. The packet loss ratio and end-to-end delay will be the QoS metrics in consideration. In the next chapter, we will apply this scheme to the end-to-end environment which is from the UMTS (WCDMA) wireless network to the DiffServ wireline network.

4.1 QoS Classes Mapping

As described in chapter 3, the UMTS QoS specifications define four classes of traffic, namely Conversational, Streaming, Interactive and Background. Different classes of traffic have different traffic specifications and QoS requirements. However, in DiffServ networks, the QoS framework is based on the implementation of different PHBs. In order to implement QoS provisioning in DiffServ networks for these four classes services, we must map the incoming UMTS traffic to the DiffServ PHBs.

Table 4.1: QoS Mapping between UMTS and DiffServ Network

UMTS QoS	Conversational	Streaming	Interactive	Background
DiffServ PHB	EF	AF1x	AF2x	AF3x
Priority	1(Highest)	2	3	4(Lowest)
Application	VoIP	Video	Web Browsing	Email, FTP

Table 4.1 shows the mapping rules between UMTS and DiffServ PHBs we set in our scheme. The most delay sensitive traffic, Conversational class, is mapped to the expedited forwarding (EF) PHB in DiffServ, which means it has the highest priority in all the traffic classes. For the Streaming class, it is a real-time traffic with a looser delay requirement, so it is mapped to AF1x PHB with the second highest priority. Non-real time traffic, Interactive and Background classes (web browsing and email, FTP etc.), have no stringent delay requirement and they are mapped to AF2x and AF3x respectively, which have the lower priorities in the network. At each router in the network, a strict priority scheduling will be used, i.e., the traffic with higher priority will be served first.

4.2 Resource Provisioning

The goal of resource provisioning is to ensure that the network has enough resources to meet the expected demand with adequate QoSs. Determination of resource required at each router for every traffic class needs the estimation of the volume of traffic that traverses each network router. Bandwidth is the most precious resource at every router in the network. We will introduce the Equivalent Bandwidth (EB) based resource allocation scheme in this chapter.

4.2.1 Equivalent Bandwidth

Over the last ten years, considerable work has been done on equivalent bandwidths. The amount of equivalent bandwidth is a value between source's mean rate and peak rate. If many sources share the same buffer, the equivalent bandwidth of each source decreases, and it would be close to the mean rate, otherwise it is close to the peak rate. Equivalent bandwidth of an ensemble of connections is usually the sum of their equivalent bandwidths. The traditional effective bandwidth is calculated based on the measurement of traffic load *A(t)* (amount of work that arrives from a source in the interval [0,t]). The asymptotics of queue length distribution in the regime of large buffers are exponential and can be characterized by two parameters, the asymptotic constant γ and asymptotic decay rate δ [24]:

$$P\{Q > B\} = \gamma e^{-\delta B} \quad , \tag{4.1}$$

where Q is the queue length in buffer and B is the buffer size. If we approximate $P\{Q > B\}$ with a cell loss rate ε, then we can arrive at equation (4.2), where γ is considered as 1.

$$\delta = -\frac{\log(\varepsilon)}{B} \quad . \tag{4.2}$$

We define the asymptotic decay rate function as

$$h_A(v) = \lim_{t\to\infty}\frac{1}{t}\log E\{\exp(vA(t))\} \quad , \tag{4.3}$$

where E is the expectation. From Large Deviation Theory and Gartner-Ellis Theorem

[49], the equivalent bandwidth of the input is given by

$$eb(\delta) = \frac{h_A(\delta)}{\delta} \ . \tag{4.4}$$

An important question is how to compute $h_A(\delta)$ for a given input traffic source. Reference [47] gives the explicit formulas of asymptotic decay rate functions for some specific traffic models, which can be used to compute the equivalent bandwidth. Equation (4.5) is the equivalent bandwidth of an exponential on-off source given by

$$eb(\delta) = \frac{r\delta - \alpha - \beta + \sqrt{(r\delta - \alpha - \beta)^2 + 4\beta r\delta}}{2\delta} \ , \tag{4.5}$$

where α is the transition rate from on to off state, β is the transition rate from off to on state, r is the transmission rate when the source is in on state, and the source rate is zero in the off state.

In the case of multiple sources, the effect of statistical multiplexing is of great significance. If the central limit theorem is applied to the traffic processes, as more sources are aggregated together, the traffic becomes more Gaussian by sharing a link with more and more traffic streams. It is appropriate to say that if sufficient traffic is aggregated, the distribution of the stationary bit rate can be rather accurately approximated by a Gaussian distribution [27,28]. The estimation of the equivalent bandwidth in this case does not take into account the buffer, i.e., bufferless model, and it generally provides an overestimation of the actual demand. This EB is given by

$$C = m + \phi\sigma \ , \quad \text{with } \phi = \sqrt{-2\ln(\varepsilon) - \ln(2\pi)} \ , \tag{4.6}$$

where m is the mean rate of aggregate sources ($m = \sum_{i=1}^{N} m_i$), σ is the standard deviation of the aggregate sources ($\sigma^2 = \sum_{i=1}^{N} \sigma_i^2$) and ε is the packet loss ratio. This equation is quite simple and depends only on the packet loss ratio, the means and variances of the bit rates of individual sources, which are directly available from the sources characteristics.

4.2.2 Equivalent Bandwidth with Priorities

There are four classes of traffic with different priorities in the scheme, and each class has different QoS requirements. Thus we need to consider the equivalent bandwidth of multiple classes with different priorities.

Assuming there are N distinct traffic classes in the system, there are K_j independent sources (j = 1, 2,…, N) producing traffic of type j that are multiplexed into a buffer of size B_j. The total service rate is c and the traffic is removed from these buffers following a static priority full service policy, i.e., traffic of type j is served before traffic of type i if $j < i$. A natural choice of the multiclass admission set would be of the form

$$K_1 eb_1 + K_2 eb_2 + \ldots + K_N eb_N < c \ , \tag{4.7}$$

where eb_j is the equivalent bandwidth of j^{th} class traffic.

However, this equation is too conservative and it does not consider the capacity gain from the multiplexing of traffic with different priorities. Berger and Whitt [26] propose equation (4.8) to replace equation (4.7) with the empty-buffer approximation and reduced-service-rate approximation. For empty-buffer approximation, we have

$$\begin{aligned} &K_1 eb_1 < c \ , \\ &K_1 eb_1^{\ 2} + K_2 eb_2 < c \ , \\ &\ldots \\ &K_1 eb_1^{\ N} + K_2 eb_2^{\ N} + \ldots + K_N eb_N < c \ , \end{aligned} \tag{4.8}$$

where $eb_j^{\,k} = \dfrac{h_{Aj}(\delta_k)}{\delta_k}$ is the equivalent bandwidth of type j sources seen by type k traffic. That is, it is the equivalent bandwidth of type j sources subject to the type k sources performance criterion. It assumes that the higher priority traffic has more stringent performance criteria than that of the lower priority traffic. Therefore $eb_j^{\,k}$ is less conservative than eb_j and improvement can be obtained. For the reduced-service-rate approximation, $eb_j^{\,k}$ is the mean rate of type j source. The main idea of equation (4.8) is that there should be multiple equivalent bandwidths associated with one type of traffic. The difference between these two approximations is that the empty-buffer approximation is still conservative while the reduced-service-rate approximation is not.

4.3 Admission Control Strategies

In section 4.1, we have mapped the four UMTS QoS classes to the different DiffServ network PHBs with their QoS requirements. According to 3GPP technique specification [16], the Conversational and Streaming classes have both the loss and delay requirements, while the Interactive and Background classes have only loss performance criteria. In this admission control scheme, the packet loss ratio guarantee will be provided by the bandwidth allocation and the end-to-end delay will be statistical guaranteed by measurement and delay bound estimation.

4.3.1 Traffic Models

Table 4.2 presents the traffic models we employ in the admission control scheme. For voice (speech) sources, the talkspurt and silence period are assumed to be exponentially distributed. A video source is modeled as a discrete state, continuous time Markov process as described in [29]. It is the summation of one high-rate (A_h) and eight low-rate (A_l) exponential on-off sources as shown in Figure 4.1, where a, b, c and d are the state transition rates of mini-sources respectively, such a model can catch video traffic's variable bit-rate characteristic. Interactive and Background traffic are modeled by Pareto-on and Pareto-off sources with different parameters.

Table 4.2: Traffic Models of UMTS QoS Classes in Simulation

UMTS QoS	Conversational	Streaming	Interactive	Background
Traffic Model	Exponential on-off	Summation of High- and Low-rate Exponential on-off sources	Pareto-on and Pareto-off	Pareto-on and Pareto-off
Application	VoIP	Video	Web Browsing	Email, FTP

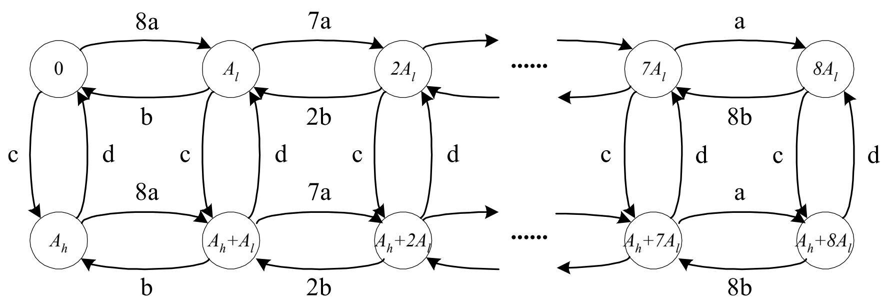


Figure 4.1: Video Source Model

4.3.2 Bandwidth Allocation

In order to provide the loss guarantee for the UMTS QoS traffic classes, it is necessary to implement the bandwidth allocation policies. The scheme in this thesis is based on the equivalent bandwidth introduced in section 4.2. There are two different types of sources (exponential on-off and Pareto on-off sources) in the four UMTS traffic classes. Their equivalent bandwidths are computed using different methods.

4.3.2.1 Exponential On-Off Source and Asymptotic Constant Estimation

For the exponential on-off sources, equation (4.5) can be used to calculate the equivalent bandwidths, and the asymptotic decay rate δ is obtained from equation (4.2). However, it does not consider the effect of asymptotic constant, γ, which is between 0 and 1. If we take γ as 1, the statistical multiplexing gains are not taken advantage of and the admission region is very conservative. The Chernoff Dominant Eigenvalue (CDE) approximation for the tail probability predicts that γ is close to the loss probability if there is no buffer, and the equation (4.2) should be rewritten as

$$\delta = -\frac{\log(\varepsilon/\gamma)}{B} \quad . \tag{4.9}$$

Then we can obtain a larger admission region. Unfortunately, γ is a function of all the parameters and the numbers of the sources of class i and the higher priority classes, so it is not easy to calculate exactly. In [25] and [30], a few methods of estimating the value of γ are suggested, but the problem of these methods is that the computation is still very complex and it is not practical for the environment where a large number of connections exist.

We propose a simple method to estimate the approximate value of the asymptotic constant of each priority class, γ_j, which is shown as follows

$$\gamma_j = \sum_{i=1}^{j} \rho_i \ , \qquad\qquad j \ = \ 1,\ 2,\ \ldots\ldots N \ , \tag{4.10}$$

where ρ_i is the utilization of class i traffic, and there are a total of N traffic classes in the system. After each updating period, the parameter γ_j will be recalculated, and the equivalent bandwidth of the new class j exponential on-off connection will be based on this value. Since γ_j is less than 1, it is less conservative than the result where γ_j equals

1. Compared to the loss probability when there is no buffer in the CDE approximation, the value of γ_j is greater, so it is more conservative than the CDE bound. Table 4.3 is the comparison of γ_j estimation using this method and with the CDE approximation. The four traffic classes parameters are shown in Table 4.5 and the number of connections of each class is 40, 4, 10, and 10, respectively. The service rate is 1 Mbps.

Table 4.3: Comparison of Asymptotic Constant Estimation

	Voice (γ_1)	Video (γ_2)	Interactive (γ_3)	Background (γ_4)
γ_j (CDE)	5.9×10^{-9}	5.6×10^{-3}	5.3×10^{-2}	3.1×10^{-1}
$\gamma_j = \sum_{i=1}^{j} \rho_j$	0.48	0.73	0.80	0.92

Table 4.4: Capacity Gain of Asymptotic Constant Estimation

Service Rate	1 Mbps	1 Mbps	5 Mbps	5 Mbps	10 Mbps	10 Mbps
Packet Loss Ratio	0.01	0.001	0.01	0.001	0.01	0.001
$\gamma_j = 1$	48	43	240	216	480	433
$\gamma_j = \sum_{i=1}^{j} \rho_j$	54	46	269	232	538	465

The value of γ_j is easy to estimate with this method, no intensive computation is needed and it is suitable when there are a large number of connections in the system. Table 4.4 shows the capacity (connection number) gain if we use the estimation method of asymptotic constant introduced above. The traffic source is voice and buffer size is 100 packets.

The problem of equation (4.5) is that when the buffer size is small, the equivalent bandwidth is very conservative and is close to the peak rate of the traffic. Thus it is more suitable for a large buffer scenario. Another approach to equivalent bandwidth is the Gaussian approximation mentioned in section 4.2.1. In this admission scheme, we choose the minimum of these two approximations as the equivalent bandwidth of voice and video sources.

4.3.2.2 Long Range Dependence Source

Long range dependent sources are sources which exhibit significant long burst periods, and the large deviation assumption may be inadequate for the type of sources that exhibit long range dependent characteristics.

In the scheme in this thesis, the Interactive and Background classes are modeled by Pareto-on and Pareto-off distribution sources. The superposition of many sources with heavy-tailed on durations is regarded as traffic models which captures the long range dependence effects of the network traffic. With the increase in the number of sources with Pareto distributed on durations, the relative contribution of each source decreases and it results in the M/Pareto model. This model assumes that the bursts arrive according to a Poisson process and the burst size follows a Pareto distribution.

Norros [33] presents an approach to obtain an approximation for the queue length distribution in an infinity buffer fed by a Fractional Brownian Motion (FBM) traffic stream and shows that the distribution of the queue length follows a Weibull distribution. The application of Norros formula to the M/Pareto traffic model can be achieved by equating the mean and the variance of the corresponding cumulated arrival processes. The equivalent bandwidth of the M/Pareto traffic model is given by

$$C = m \cdot \left(1 + x(H) \cdot (-2\ln\varepsilon)^{\frac{1}{2H}} \cdot \left(\frac{B}{b}\right)^{\frac{H-1}{H}} \cdot \left(\frac{r}{m}\right)^{\frac{2H-1}{2H}} \right), \tag{4.11}$$

where $x(H) = 2^{\frac{1-H}{H}} (3-2H)^{-\frac{3-2H}{2H}} \cdot H^{\frac{2H-1}{2H}} \cdot (1-H)^{\frac{2-2H}{H}} \cdot (2H-1)^{-\frac{1}{2H}}$. In the equation, *m* is the total mean rate, *b* is the mean burst size, *r* is the peak rate of burst, *B* is the buffer size, and ε is the loss ratio. The Hurst parameter *H* of the M/Pareto model has the following relationship with the shape parameter, ω, of the Pareto distribution ($P(x) = \frac{\omega\theta^{\omega}}{x^{\omega+1}}$) for the on period:

$$H = \frac{3-\omega}{2}\ , \qquad \frac{1}{2} \leq H < 1\ . \tag{4.12}$$

With the aggregation of more and more traffic streams, the traffic becomes more Gaussian, or more formally, weakly converges to a Gaussian process [31,34]. Bodamer and Charzinski [32] propose a simple approach to the M/Pareto traffic equivalent bandwidth which combines the Gaussian approximation and equation (4.11) to achieve a less conservative admission region. We employ this approximation to calculate the equivalent bandwidth of Interactive and Background classes traffic using the Newton-Raphson numerical method.

4.3.3 Statistical Delay Guarantee

Since Conversational and Streaming classes are real-time traffic, the end-to-end packet delay in the network is also an important QoS metric to be considered in addition to the packet loss ratio.

For Conversational class, the delay distribution is similar with its workload distribution, because it is the highest priority class in the system. However, the delay of the lower priority class, Video Streaming, is different from its workload due to the effect of the higher priority class. It is difficult to get the exact low priority class delay distribution, although reference [48] gives some tail asymptotics for the waiting times. The problem is that it is not easy to obtain closed-form expressions and simulation results show that the asymptote is not accurate.

In the scheme in this thesis, we use the simple approximation for the waiting time tail probabilities in a multiclass system [35] to estimate the delay bound at each router in the network. The expression for the waiting time tail probability is given by

$$P(W_i > x) \approx \rho e^{-\rho x / \overline{w_i}} \quad , \tag{4.13}$$

where W_i is the waiting time of class i packet, $\overline{W_i}$ is the average waiting time of class i packets, and ρ is the system utilization. The routers measure and update the average waiting times of the Conversational and Streaming classes periodically, and then calculate the queuing delay bounds according to the statistical guarantee requirements. The total system delay bounds with the addition of fixed service times of each corresponding class packet are then computed.

4.4 Single-Hop Scenario

In this section, a single edge router model is investigated. Four UMTS QoS traffic classes arrive to the system from the end users and each traffic class has its own buffer. The router serves the packet with strict priority scheduling, i.e., transmit the packet with the higher priority first.

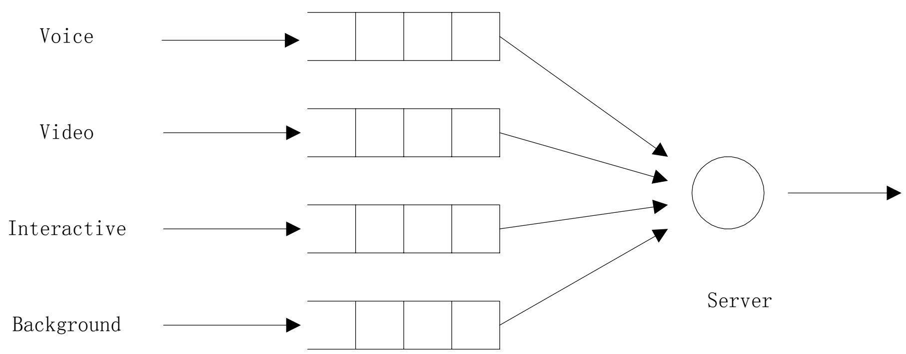


Figure 4.2: Single-Hop Scenario

Table 4.5 gives the traffic and server parameters in the simulation, where μ is the mean value of exponential distribution and r is the peak rate during the on period.

Table 4.5: Simulation Parameters

Voice	On period	μ = 1 s, r = 30000 bps
	Off period	μ = 1.5 s
	Holding Time	Exponential, μ = 300 s
	Packet Size	600 bits
Video	On period	μ = 1.5 s, r = 30000 bps (High rate), μ = 0.42 s, r = 15000 bps (Low rate)
	Off period	μ = 1.5 s (High rate), μ = 0.66 s (Low rate)
	Holding Time	Exponential, μ = 300 s
	Packet Size	900 bits
Interactive	On period	Pareto, Location = 0.1455 s, Shape = 1.1, r = 60000 bps
	Off period	Pareto, Location = 1.0909 s, Shape = 1.1
	Holding Time	Lognormal, Median = 300 s, Shape = 2.5
	Packet Size	1200 bits
Background	On period	Pareto, Location = 0.268 s, Shape = 1.1, r = 120000 bps
	Off period	Pareto, Location = 2.3273 s, Shape = 1.1
	Holding Time	Lognormal, Median = 5 s, Shape = 2.5
	Packet Size	2400 bits
Server	Scheduling	Strict Priority
	Service Rate	10 Mbps
Connection	Inter-Arrival	Poisson Arrival, Exponential Distribution

In the admission control of a single-hop case, only the equivalent bandwidth will be considered, and we will discuss the buffer and bandwidth management for the different priority classes in the system. There are two multiclass bandwidth allocation schemes to be studied in this simulation: (A) a simple summation of all the equivalent bandwidth of four classes traffics, and (B) a reduced-service-rate approximation introduced in section 4.2.2. The reason that empty-buffer approximation is not employed is that the higher priority traffic classes, i.e., the real-time traffic (Conversational and Streaming) have looser packet loss ratio requirements than those of the non-real-time traffic (Interactive and Background) in UMTS (WCDMA).

4.4.1 Buffer Management

Since the four UMTS QoS traffic classes in the system have different priorities in the service order, the buffer sizes required for each class have a great impact in the network.

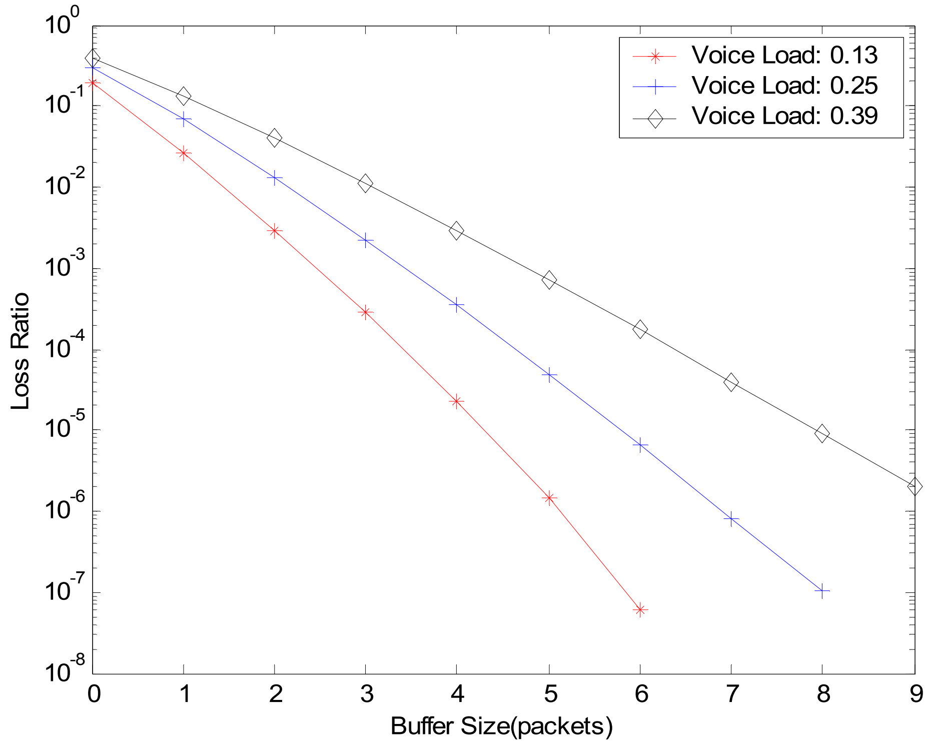


Figure 4.3: Voice Packet Loss Ratio vs. Buffer Size

Figure 4.3 and 4.4 show the voice and video packet loss ratio versus their corresponding buffer size, respectively, where the multiclass bandwidth allocation scheme B, i.e., the reduced-service-rate approximation is used. Since the buffer size is small and the number of voice and video connections is large, the equivalent bandwidth from Gaussian approximation will be the minimum of the two methods mentioned for exponential on-off sources.

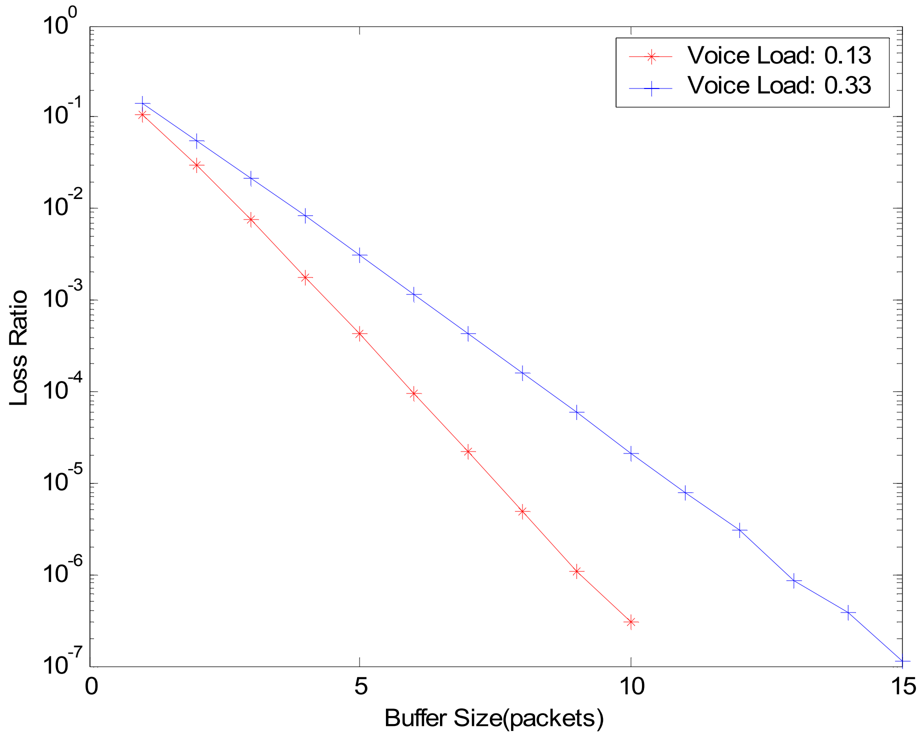


Figure 4.4: Video Packet Loss Ratio vs. Buffer Size (Video Load: 0.33)

In theory, the Gaussian approximation should be able to provide the desired loss ratio even with no buffer. In fact, from Figure 4.3, we can see that the loss ratio for voice packet is almost 20% without buffer, and is 3% with just one packet buffer even when the voice load is very light (0.13). This is due to the collisions of packets arriving from different input links in very small time intervals. The equivalent bandwidth treats the traffic as a continuous fluid, and the loss ratio computed represents the overflow probability in case that the sum of incoming flows exceeding the service rate. However, the real traffic is not continuous. They arrive from every input link with their individual packets. Even with a high service rate, the router can handle only one packet at each time. If there is no buffer and two packets arrive within the interval smaller than one packet service time, one of them will be dropped. This will be even worse if the number of input links is very large, such as at the edge routers of a DiffServ network, which received the incoming packets from many end hosts. In the core of networks, the effect of the collisions will decrease due to the small number of input links and the high service rate at the core routers.

Since the high service rate is not sufficient to achieve the low loss requirement, the buffer size needs to be increased. From Figure 4.3, a four packets buffer size is enough to guarantee the packet loss ratio of below 10^{-2} for voice traffic even for high load (0.39). Actually, the traffic volume of Expedited Forwarding PHB in the DiffServ network cannot be so high and has a limit. Otherwise, it will impose a great effect on the lower priority classes. For the video traffic sources, the loss ratio performance is similar to the voice traffic. The difference is that the voice load has a great effect on the video packet loss. In Figure 4.4, although the two curves have the same video load of 0.33, the loss ratio of the video traffic with a voice load of 0.32 is much higher than the one with a voice load of 0.13. It is reasonable since the higher priority classes have the privilege of being served first. Even with a high voice load, it only needs a buffer size of five packets for the video traffic to obtain a loss ratio of below 10^{-2}.

For the third priority class, Interactive class, the loads of voice and video traffic play an important role on its loss performance, since these are higher priority classes. As Interactive applications are data traffic sources, they have higher packet loss requirement than that for the real-time traffic. Thus the loss criterion is set to 10^{-3} for data traffic at each router in the simulation. Figure 4.5 shows the packet loss ratio of Interactive class traffic versus its buffer size.

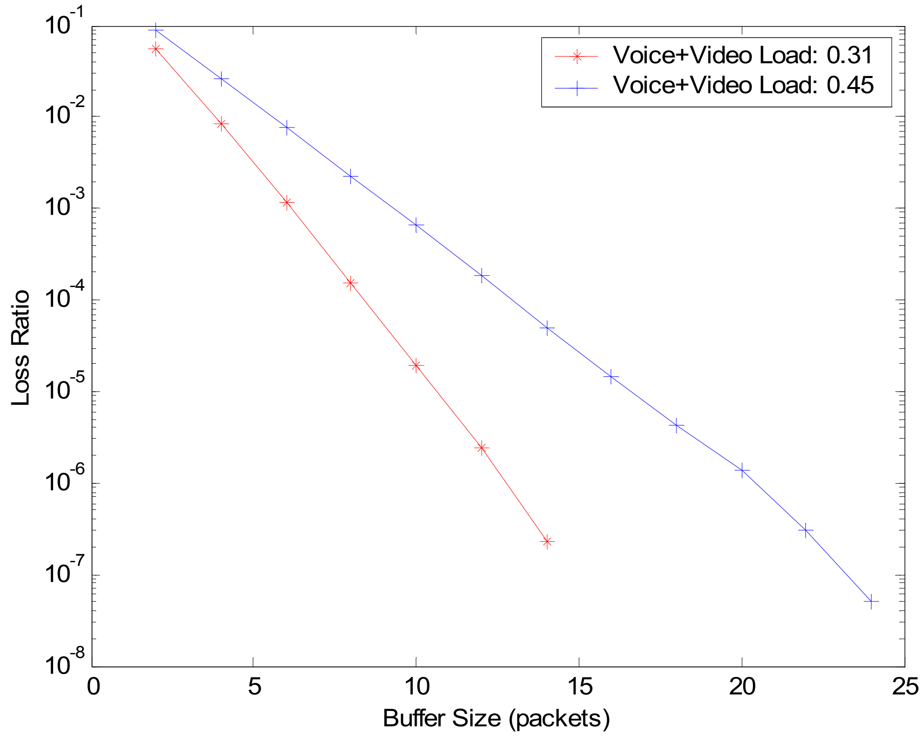


Figure 4.5: Interactive Packet Loss Ratio vs. Buffer Size (Interactive Load: 0.33)

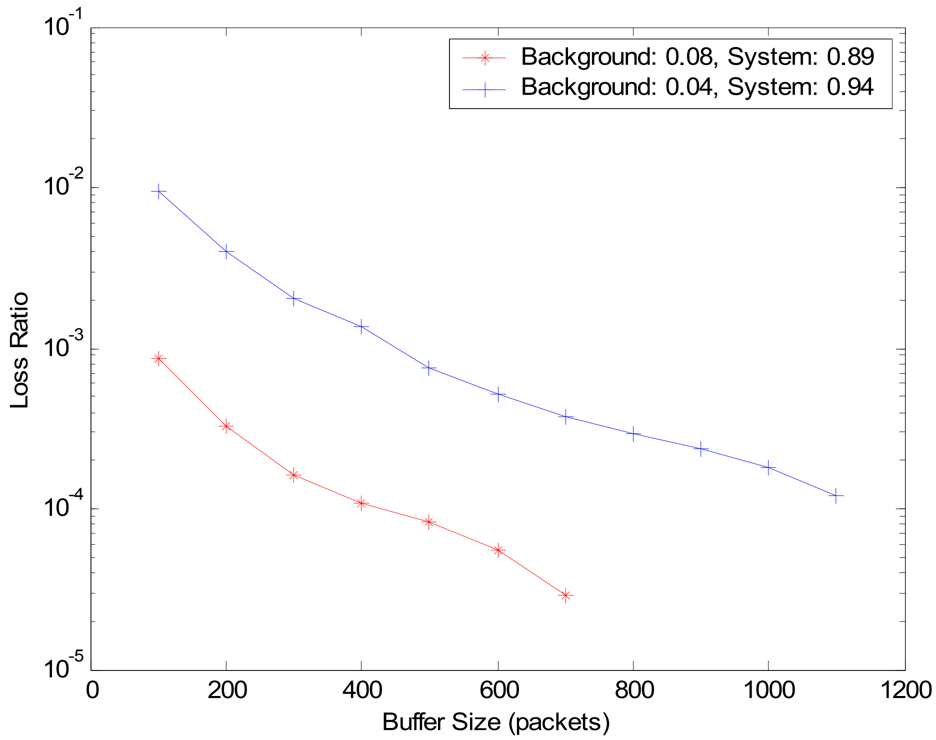


Figure 4.6: Background Packet Loss Ratio vs. Buffer Size

Background traffic is the lowest priority class in the system. Figure 4.6 shows its packet loss versus buffer size in the single-hop simulation. We can see that the system load has a great effect on its loss performance, especially when it is close to 1. It is reasonable since the Background traffic can only use the bandwidth in the case when there are no other traffic classes. In order to guarantee the loss ratio to be below 10^{-3}, Background class needs a much larger buffer size than the other three classes in the high load situation, and the decrease rate of packet loss ratio with buffer size increase is much slower.

In Figure 4.7, the packet loss ratios of voice and video packets with their corresponding queuing delay are presented. We can see that in order to achieve a low loss ratio, we can increase the buffer size. However, the increase of the queuing delay is so small that it can almost be neglected.

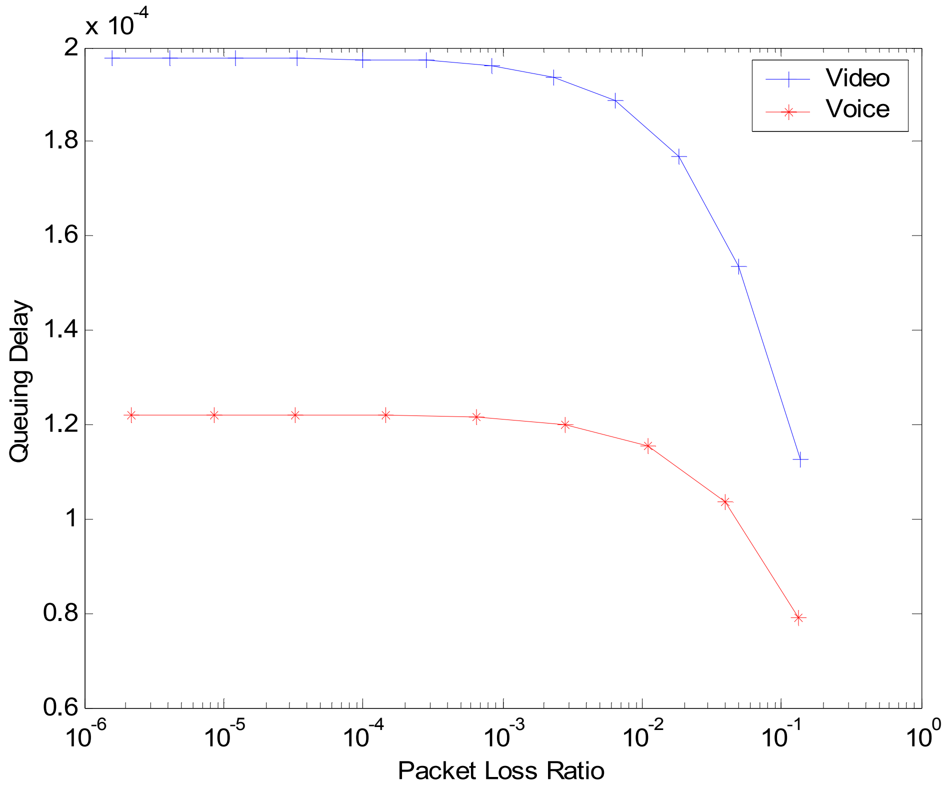


Figure 4.7: Voice and Video Packet Loss Ratio vs. Queuing Delay

The simulation results suggest that for real-time traffic, even the EF traffic, buffering is still needed to achieve a low loss requirement, especially at the router with many input links, such as the edge routers of a DiffServ network. However, the buffer size needed at the routers is small, only several packets size can provide the 10^{-2} loss guarantee in the simulation. This is because that voice and video are the two highest priority classes in the network and especially the voice traffic (Conversational) has the privilege to use most of the bandwidth resources. As a result, the buffer occupancies for these two classes are very low. Large buffer does not appear to impose too much an increase to the delay, but can reduce the packet loss ratio significantly.

For non-real-time traffic, Interactive and Background classes, they have more stringent loss requirements. Thus the buffer sizes needed are larger. The total load of each class and its higher priority classes has a great effect on the loss performance for this class. Background traffic is the lowest priority class in the system and its loss performance is very sensitive to the system load, especially when the system load is close to 1. It will suffer great packet loss, so it needs a much larger buffer size. The results in this section can provide helpful information for the multiclass system design.

4.4.2 Multiclass Bandwidth Management

Two multiclass bandwidth allocation schemes are examined in the simulation: (A) a simple summation of all the equivalent bandwidths of the four traffic classes (conservative method), and (B) a reduced-service-rate approximation (non-conservative method).

Obviously, scheme A is more conservative than that of scheme B and can result in a lower system utilization. Table 4.6 shows three set of results for these two schemes under different user arrival rates. In the simulation, the buffer size of each class is 5, 6, 30 and 600 packets respectively and the loss ratio requirement is 10^{-2} for voice and video, and 10^{-3} for Interactive and Background traffic. In order to make the system reach nearly the full utilization state, we set the user arrival rates to relatively high values, which are shown in Table 4.6.

Table 4.6: Single-Hop System Utilization

	Simulation I			Simulation II			Simulation III		
	Mean Inter-Arrival Time	Util (A)	Util (B)	Mean Inter-Arrival Time	Util (A)	Util (B)	Mean Inter-Arrival Time	Util (A)	Util (B)
Voice	1.5 s	0.085	0.093	0.5 s	0.35	0.37	0.5 s	0.369	0.41
Video	0.1 s	0.275	0.42	0.3 s	0.085	0.206	0.3 s	0.195	0.31
Interactive	10 s	0.269	0.299	10 s	0.306	0.313	20 s	0.17	0.183
Background	0.5 s	0.091	0.106	1.5 s	0.029	0.046	1.5 s	0.036	0.038
Total	N/A	0.72	0.919	N/A	0.769	0.932	N/A	0.768	0.94

From numerical results in the table, we can see that scheme B can achieve a much higher system utilization than that of scheme A. There is nearly twenty percent increase in the above three simulations. The reason that the video traffic sources achieve the largest utilization increase is that we set a relatively higher arrival rate for the video users compared to the other traffic classes. Since one video user has a much higher equivalent bandwidth compared to voice, Interactive and Background traffic, if they arrive at the same rate, the video user request is more prone to be rejected and squeezed out of the system. Of course, we can limit the traffic volume of each class in the network to prevent such unfairness. The equivalent bandwidth of long range dependence sources is more conservative than that of the exponential on-off sources due to their long period bursts. Thus the higher the percentage of long range dependence traffic volume, the lower the system utilization.

In the simulations of Table 4.6, the experienced packet loss ratios are all within their respective bounds. Actually for scheme A in the large buffer situation, no packet loss is observed in some simulations since it is conservative and the utilization is really low. Scheme B is a non-conservative approach. Thus it could provide higher system utilization, but violate the loss performance criteria in some cases.

The packet loss ratio performance of the scheme A and B are presented in Table 4.7. In each simulation scenario, the mean inter-arrival time (μ) of each class user is the same for the two schemes, but the buffer size is set to different values.

Table 4.7: Single-Hop Packet Loss Ratio

		Voice			Video			Interactive			Background			Total
		Buffer (pkts)	Util	Loss Ratio	Buffer (pkts)	Util	Loss Ratio	Buffer (pkts)	Util	Loss Ratio	Buffer (pkts)	Util	Loss Ratio	Util
I	μ	0.5 s			0.1 s			20 s			0.5 s			N/A
	A	5	0.3	8.1×10^{-4}	5	0.2	4.9×10^{-4}	6	0.15	1.3×10^{-4}	6	0.09	2.8×10^{-6}	0.72
	B	5	0.31	1.1×10^{-3}	6	0.35	5.7×10^{-3}	20	0.15	4.6×10^{-6}	300	0.1	3.8×10^{-4}	0.91
II	μ	0.5 s			0.2 s			20 s			0.3 s			N/A
	A	5	0.29	9.8×10^{-4}	5	0.1	3.7×10^{-5}	6	0.13	1.3×10^{-5}	6	0.16	3.6×10^{-5}	0.69
	B	5	0.3	1.6×10^{-3}	6	0.23	5.1×10^{-4}	10	0.16	3.9×10^{-5}	100	0.17	8.3×10^{-4}	0.87

As described before, the system utilization of scheme A is low, so the buffer size it needs to guarantee the loss ratio bound is also small. From the data in the above table, it is shown that a buffer size of less than ten packets is enough to provide the loss requirement in scheme A. For scheme B, the buffer sizes needed for the two highest priority classes, voice and video, does not have much difference as compared to those in scheme A. However, the two lower priority classes, Interactive and Background, need larger buffers to keep low loss ratios. In particular, for the Background traffic, the buffer size is much larger since it has the lowest priority and the system load is high. This conforms to the simulation results obtained in section 4.4.1.

Scheme A is a conservative approach and it provides a lower bound for the final admissible set, while scheme B (reduced-service-rate approximation) provides the upper bound and it is a non-conservative approach. If we want to provide hard guarantee on the loss ratio, we can select scheme A. However, we have to accept the underutilization of the system. On the contrary, if scheme B is used, due to its non-conservativeness, the loss ratio will be violated in some cases. In order to achieve acceptable system utilization while providing the loss ratio guarantee, we can choose another approach, scheme C. In this scheme, we use the reduced-service-rate approximation but we do not allocate all the link bandwidth. Instead, we leave a few percentage of capacity to prevent too high a load which causes too large a loss ratio.

Table 4.8 presents the packet loss ratio comparison between schemes B and C. The buffer size is set to 5, 6, 15, and 100 packets for the four traffic classes, respectively, and 5% of the link bandwidth is reserved in scheme C. From the numerical results, we can see that in some cases, scheme B results in too high a system utilization and violates the loss bound, while scheme C can avoid such a situation and satisfy the packet loss guarantee, especially for the Background traffic. The problem is how to decide the percentage of the bandwidth to be reserved. In general, five percent is enough to bound the packet loss ratio in this simulation scenario.

Table 4.8: Packet Loss Ratio Comparison between Scheme B and C

		Voice		Video		Interactive		Background		Total
		Util	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util
I	μ	1.5 s		0.1 s		10 s		1.5 s		N/A
	B	0.11	6.5×10^{-6}	0.48	3.6×10^{-3}	0.3	4.3×10^{-3}	0.05	3.1×10^{-2}	0.95
	C	0.1	2.9×10^{-6}	0.47	3.3×10^{-3}	0.27	4.1×10^{-4}	0.04	1.1×10^{-3}	0.88
II	μ	1.5 s		0.3 s		10 s		1.5 s		N/A
	B	0.12	1.9×10^{-5}	0.37	9.8×10^{-4}	0.36	1.3×10^{-3}	0.06	2.1×10^{-2}	0.92
	C	0.13	1.3×10^{-5}	0.37	7.5×10^{-4}	0.33	3.1×10^{-4}	0.05	8.1×10^{-4}	0.88

4.4.3 Admission Region

In practice, since the multiclass bandwidth allocation is only related to the traffic characteristics and the buffer size of each class, we can directly obtain the admission region of the multiclass system. In the actual application, the only job for the Bandwidth Broker to do when a new connection request arrives is to check the database for the admission region. If the combination of the numbers of users from different services is in the table, it accepts the request. Otherwise it rejects the request.

Figure 4.8 presents two examples of the admission region of scheme A, where the number of Background users is set to 3 or 13, and the service rate is 1 Mbps. Figure 4.9 shows the admission region of scheme B, where we can see that it is much larger as compared to scheme A. The regions with different number of Background users (3 or 13) in scheme B do not have much difference. It is reasonable that the mean rate of the higher priority traffic class is employed as it is the equivalent bandwidth seen by the lower priority class in the scheme.

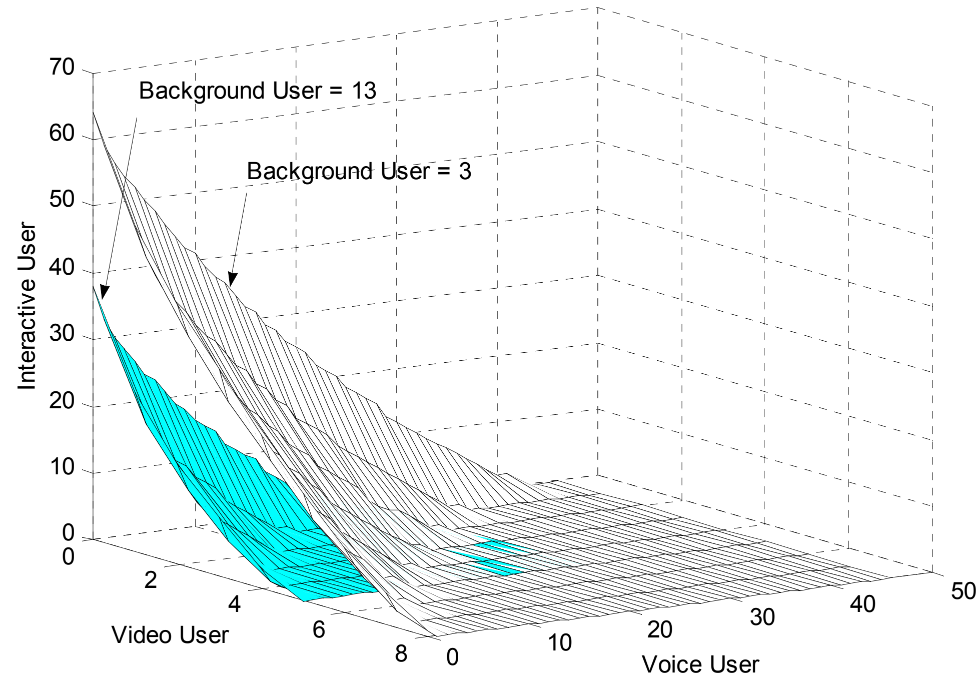


Figure 4.8: Admission Region Examples of Scheme A

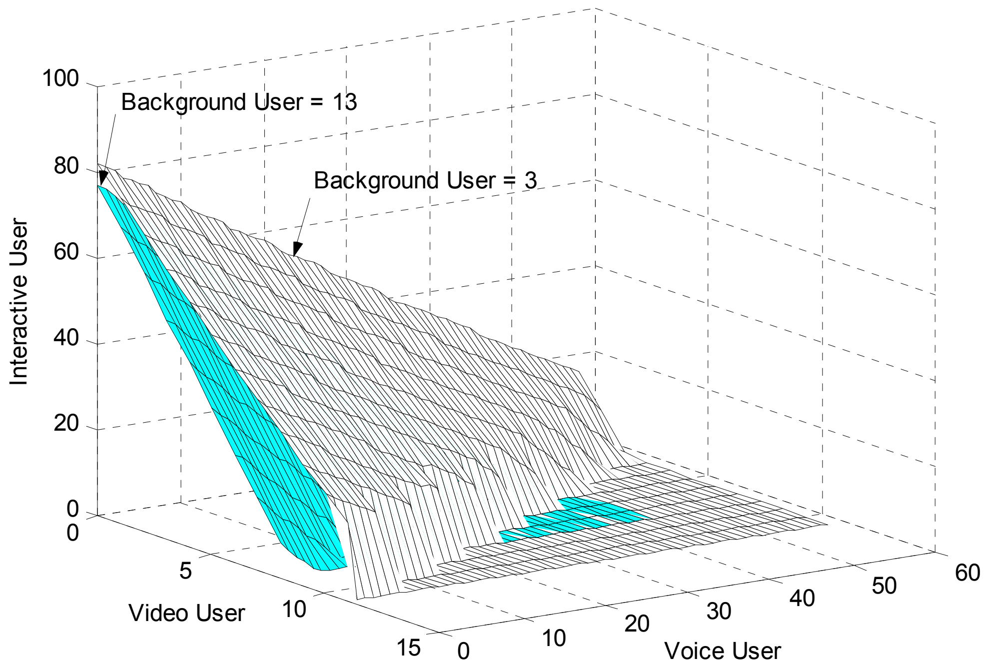


Figure 4.9: Admission Region Examples of Scheme B

4.5 Multi-Hop Scenario

We will investigate multi-hop cases in this section. The packet loss ratio and delay will both be the admission control criteria in the simulation. The simulations are performed using OPNET, and the network topology is shown in Figure 4.10.

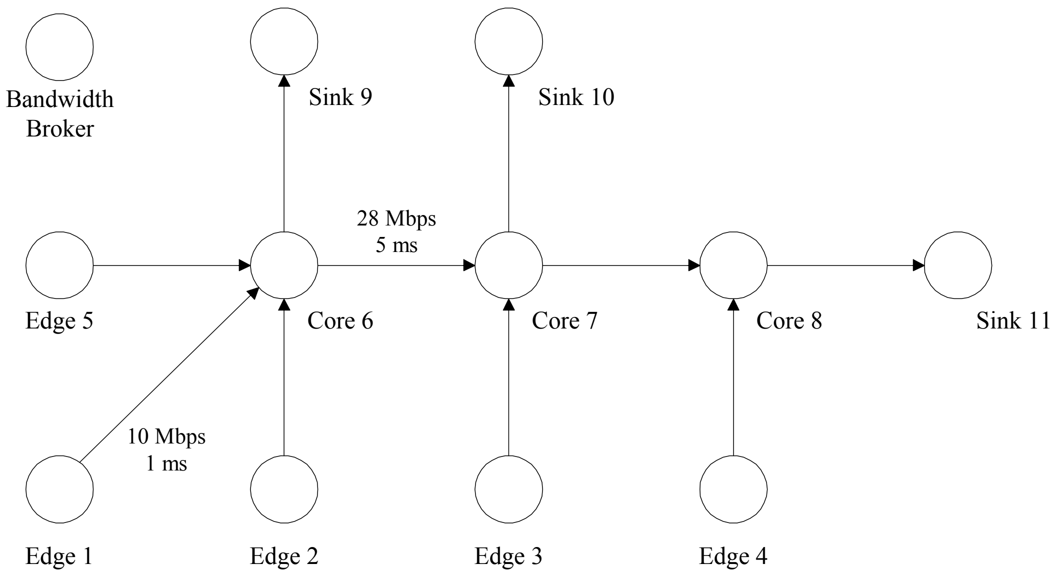


Figure 4.10: Multi-Hop Simulation Topology

There are five edge routers (Edge 1 to Edge 5) and three core routers (Core 6 to Core 8) in the network. The packets arrive to the edge routes directly and are transmitted to the core routers, and then received at the sinks (Egress routers). Table 4.9 gives the source-destination pairs in the simulation.

Table 4.9: Source-Destination Pairs

Source	Edge 1	Edge 2	Edge 3	Edge 4	Edge 5
Destination I	Sink 10	Sink 10	Sink 10	N/A	Sink 9
Destination II	Sink 11	Sink 11	Sink 11	Sink 11	N/A

The queuing buffer model and scheduler in the edge and core routes are the same as with the single-hop node case in Figure 4.2. The service rate and output link rate is 10 Mbps for edge routers and 28 Mbps for core routers. The propagation delay is 1 ms and 5 ms, respectively. There is Bandwidth Broker in the network which is responsible for the admission control and resource allocation. The edge and core routers collect the measurement results periodically, and the admission decision is made based on these information and the network status.

To maintain consistency and comparability with the previous results, the loss target of each router is the same as with the single-hop simulation, i.e., 10^{-2} for real-time traffic (Voice and Video) and 10^{-3} for non-real-time traffic (Interactive and Background). Roughly speaking, the network loss probability for a connection is approximately the loss probability of a typical router multiplied by the number of routers through which a connection traverses assuming a low packet loss probability at each router. For the packet delay, we set a 99 percent delay guarantee at each router. This means that the packets which exceed delay QoS requirement should be less than 1 percent, and the end-to-end delay bound violation probability is also approximately the sum of all the violation probabilities at each router along the connection path.

4.5.1 Admission Control Algorithm

For every new connection request, the two conditions of Equivalent Bandwidth and delay constraints must both be satisfied on each link of the new connection path before the network admits it. If either of the conditions can not be satisfied, the admission request will be rejected. The final admission control decision is only made at the Bandwidth Broker.

- Equivalent Bandwidth: Scheme B, i.e., the reduced-service-rate approximation, introduced in the single-hop simulation is employed as the bandwidth allocation algorithm.
- Delay Requirement: The statistic end-to-end delay bound calculated using equation (4.13) for the new connection should be less than its delay requirement and the statistical guarantee percentage should also be satisfied. When a new connection (Voice, Conversational class) arrives, the end-to-end delay of every existing voice and video connection on each link of new connection's path will be checked. It should be less than its delay requirement. When a new connection (Video, Streaming class) arrives, the end-to-end delay of all existing video connections on each link of new connection's path will be checked, and it should be less than its delay requirement. On the other hand, when a new connection (Interactive and Background classes) arrives, the delay of existing voice and video connections on each link of the new connection's path will not be checked.

4.5.2 Simulation

Table 4.10 presents the parameters used in the multi-hop simulation, which includes the buffer size, user arrival rates and end-to-end delay bound. As the core routers have few input links and the service rate is high, the possibility of having packet collision is low. Thus large buffer sizes are not necessary and we set a relatively small value for them in the core routers as compared to the edge routers. In the settings of end-to-end delay bounds, they are decided according to the connection path length, i.e., the more hops, the longer the delay bound and the lower the statistical guarantee percentage. The total simulated time is 15000 s, and the data is recorded from 5600 s onwards. The earlier stage is the warm-up period.

Table 4.10: Simulation Parameters

<table>
<tr><td colspan="2"></td><td colspan="2">Voice</td><td colspan="2">Video</td><td colspan="2">Interactive</td><td colspan="2">Background</td></tr>
<tr><td rowspan="2">Buffer (pkts)</td><td>Edge</td><td colspan="2">4</td><td colspan="2">7</td><td colspan="2">10</td><td colspan="2">50</td></tr>
<tr><td>Core</td><td colspan="2">4</td><td colspan="2">5</td><td colspan="2">6</td><td colspan="2">10</td></tr>
<tr><td colspan="2">Mean Inter-arrival Time</td><td colspan="2">0.5 s</td><td colspan="2">0.1 s</td><td colspan="2">20 s</td><td colspan="2">0.5 s</td></tr>
<tr><td colspan="10"></td></tr>
<tr><td colspan="2">Src-Dest Pair</td><td>1-10</td><td>1-11</td><td>2-10</td><td>2-11</td><td>3-10</td><td>3-11</td><td>4-11</td><td>5-9</td></tr>
<tr><td colspan="2">Voice Delay (ms)</td><td>11.7</td><td>16.85</td><td>11.7</td><td>16.85</td><td>6.55</td><td>11.7</td><td>6.55</td><td>6.55</td></tr>
<tr><td colspan="2">Video Delay (ms)</td><td>12.25</td><td>17.45</td><td>12.25</td><td>17.45</td><td>7.1</td><td>12.25</td><td>7.1</td><td>7.1</td></tr>
</table>

4.5.2.1 Packet Loss Ratio and System Utilization

Table 4.11 gives the packet loss ratio and utilization experienced in the simulation. From the numerical results, we observe that the packet loss ratios of the four QoS classes at each router are well-guaranteed.

At the core routers, the number of connections is much larger, so the individual connection equivalent bandwidth is less conservative than that of the edge routers. As a result, the bandwidth allocation efficiency is also higher. Compared with the result of the single-hop scenario which does not have delay requirement, the utilizations of routers in the multi-hop simulation do not reach their full capacities. This is because the delay criterion of the admission control in this case is more stringent than the equivalent bandwidth allocation criteria. It will reject the new connection request first and thus reduce the system utilization. It is clearer if we have a look at the voice and video utilization of Edge 1 and Edge 2 routers. They are lower than those of Edges 3, 4 and 5. The connections from Edges 1 and 2 to the corresponding receivers must traverse more links. The exhaustion of the available bandwidth or any existing connection's delay bound violation on any link of the path will cause admission failure. Thus the request is more prone to be rejected than the connections originating from the other edge routers, which results in lower voice and video traffic utilization.

Table 4.11: Packet Loss Ratio and Utilization

	Voice		Video		Interactive		Background		Total
	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Util
Edge 1	3.7×10^{-4}	0.128	1.6×10^{-4}	0.331	5.5×10^{-6}	0.163	3.1×10^{-5}	0.191	0.81
Edge 2	3.5×10^{-4}	0.128	1.4×10^{-4}	0.326	7.3×10^{-6}	0.18	4.3×10^{-5}	0.19	0.82
Edge 3	9.2×10^{-4}	0.172	5.8×10^{-4}	0.381	2.2×10^{-5}	0.156	3.8×10^{-4}	0.155	0.86
Edge 4	8.1×10^{-4}	0.167	5.4×10^{-4}	0.377	9.7×10^{-5}	0.177	8.3×10^{-4}	0.155	0.88
Edge 5	1.8×10^{-3}	0.214	8.1×10^{-4}	0.373	1.1×10^{-4}	0.162	9.6×10^{-4}	0.138	0.89
Core 6	2.9×10^{-4}	0.168	9.6×10^{-5}	0.368	7.2×10^{-6}	0.18	3.1×10^{-4}	0.185	0.9
Core 7	6.5×10^{-4}	0.153	2.3×10^{-4}	0.371	6.8×10^{-6}	0.178	9.1×10^{-5}	0.191	0.89
Core 8	8.5×10^{-4}	0.159	6.6×10^{-4}	0.365	1.5×10^{-4}	0.18	8.9×10^{-5}	0.182	0.89

4.5.2.2 End-to-End Packet Delay Guarantee

Besides the packet loss ratio, the end-to-end packet delay is also an important QoS metric to be guaranteed in a multi-hop scenario. Figures 4.11 to 4.26 present the end-to-end voice and video packet delay distribution, which are given by the different source-destination pairs. Since the packet delay at each router is set at the 99th percentile guarantee, the more hops on the path, the lower the end-to-end delay guarantee percentile.

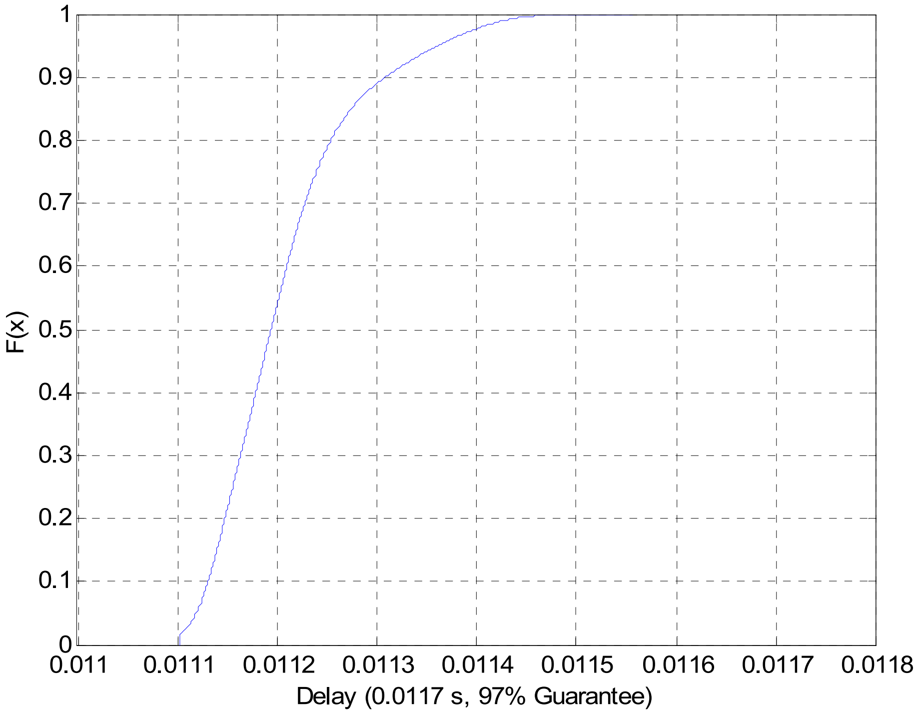


Figure 4.11: Voice Packet Delay Distribution (Edge 1 – Sink 10)

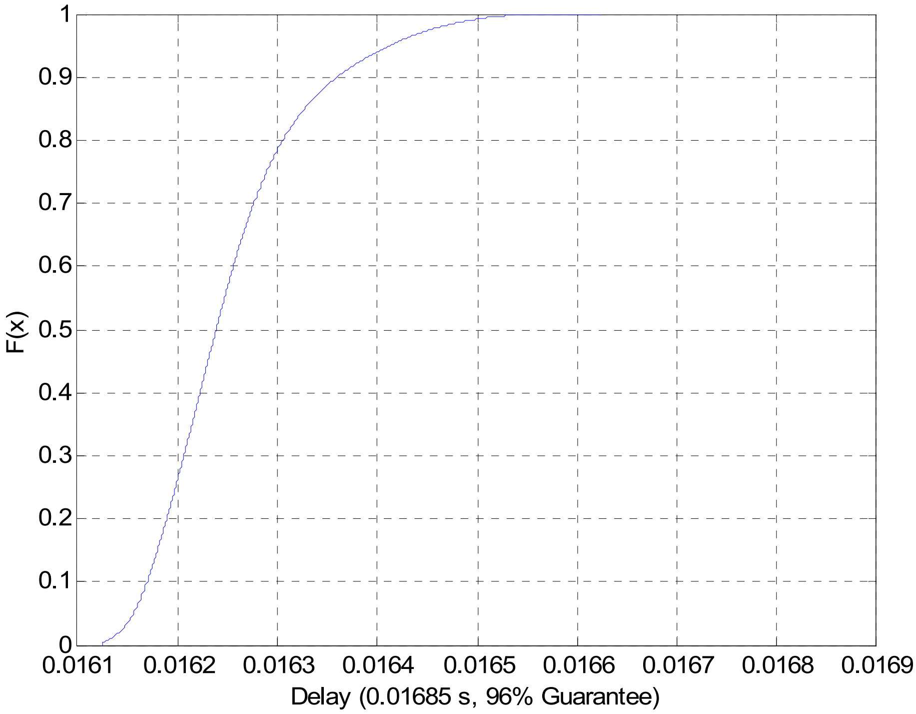


Figure 4.12: Voice Packet Delay Distribution (Edge 1 – Sink 11)

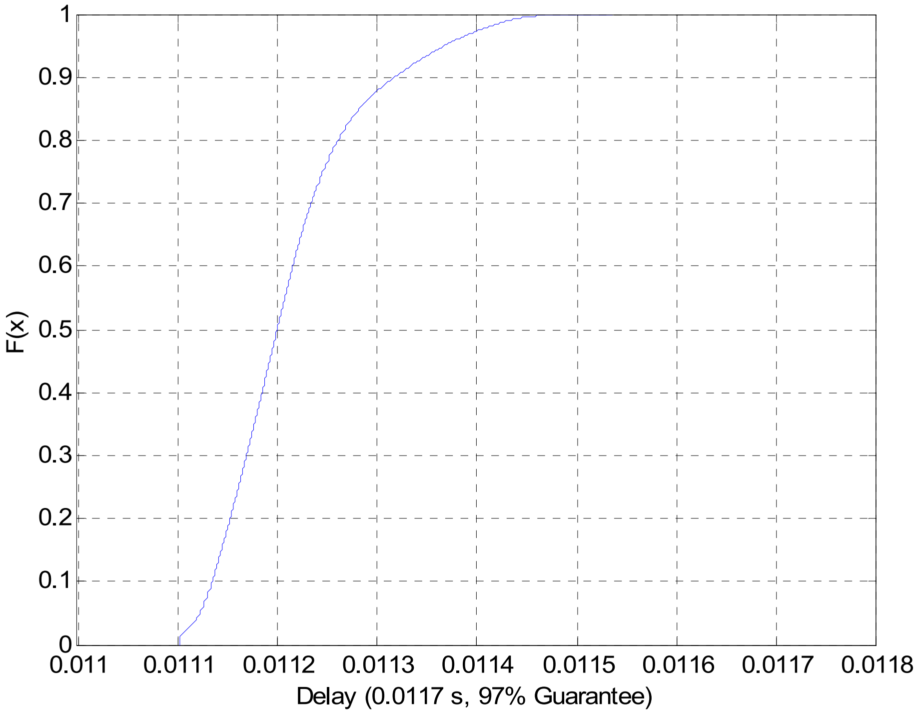


Figure 4.13: Voice Packet Delay Distribution (Edge 2 – Sink 10)

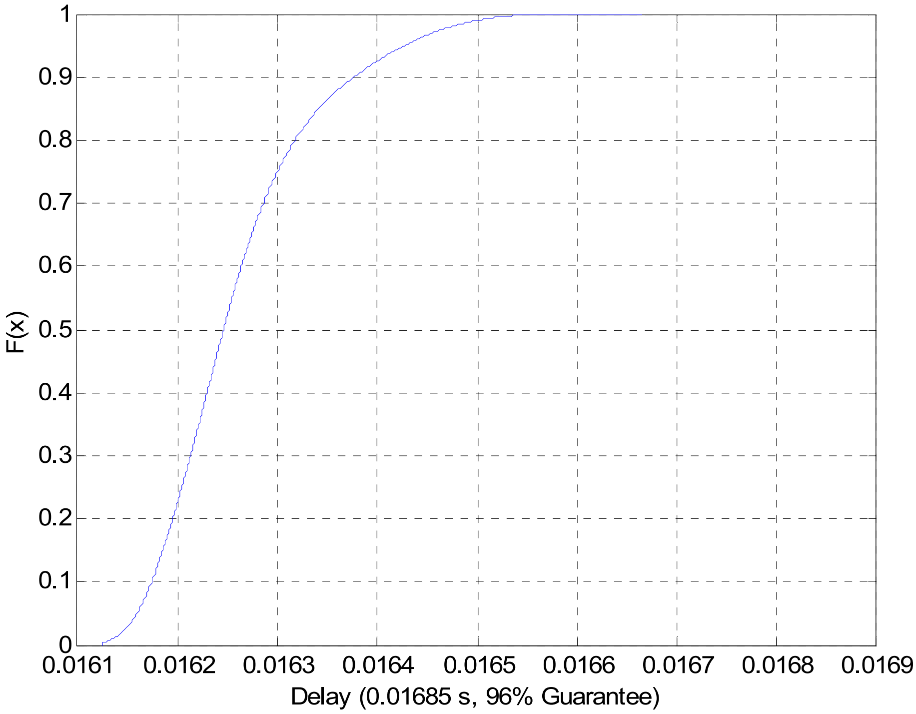


Figure 4.14: Voice Packet Delay Distribution (Edge 2 – Sink 11)

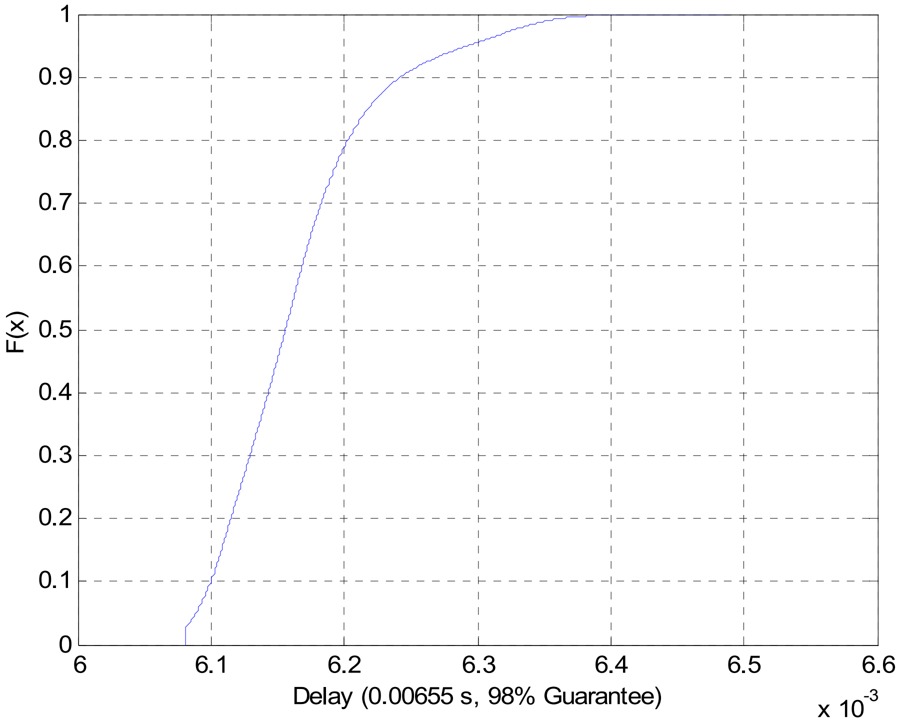


Figure 4.15: Voice Packet Delay Distribution (Edge 3 – Sink 10)

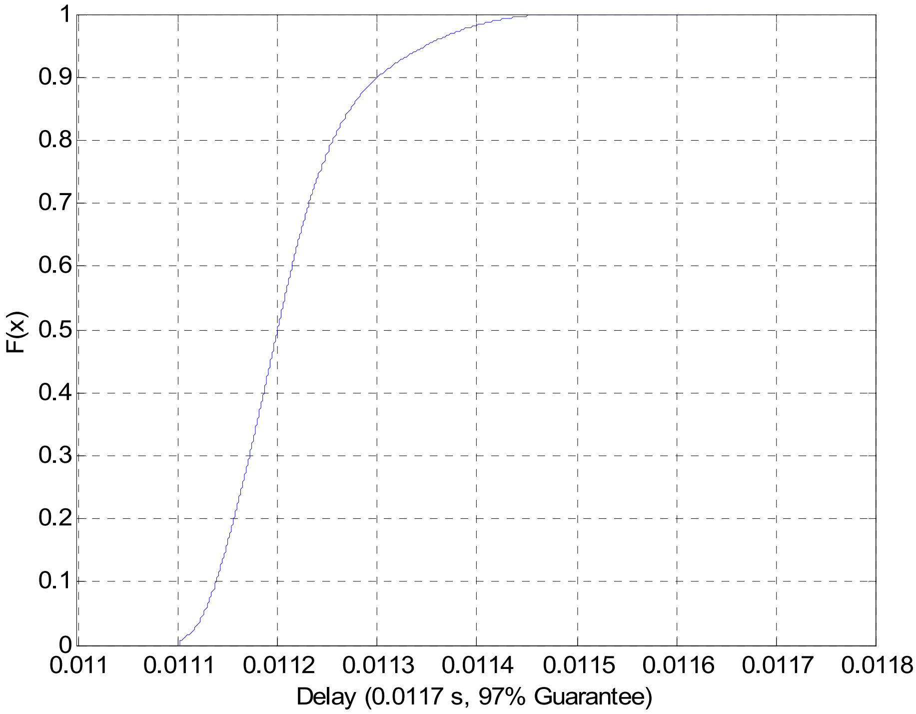


Figure 4.16: Voice Packet Delay Distribution (Edge 3 – Sink 11)

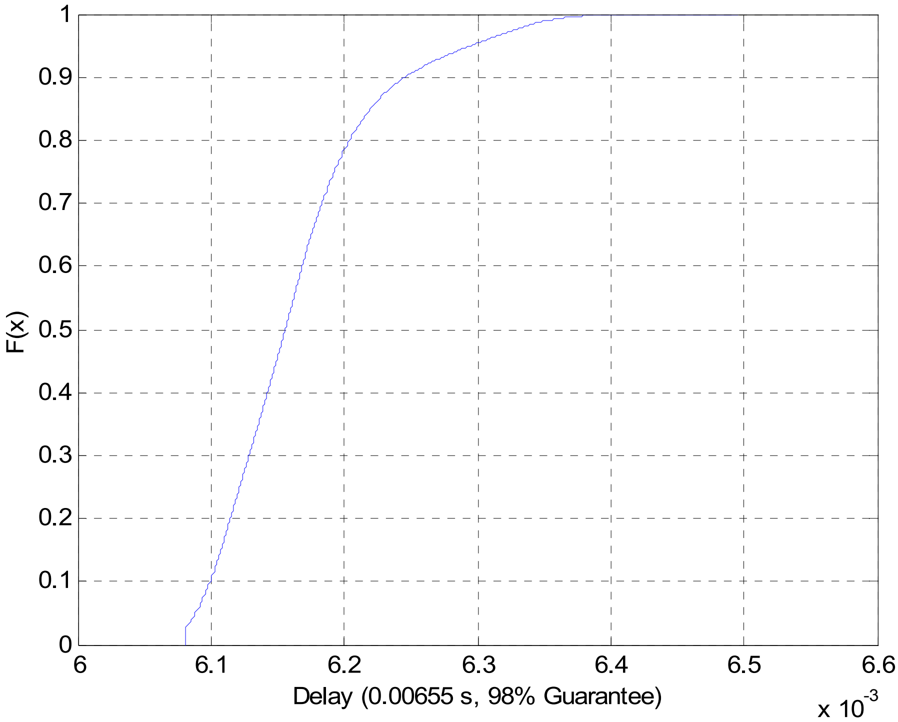


Figure 4.17: Voice Packet Delay Distribution (Edge 4 – Sink 11)

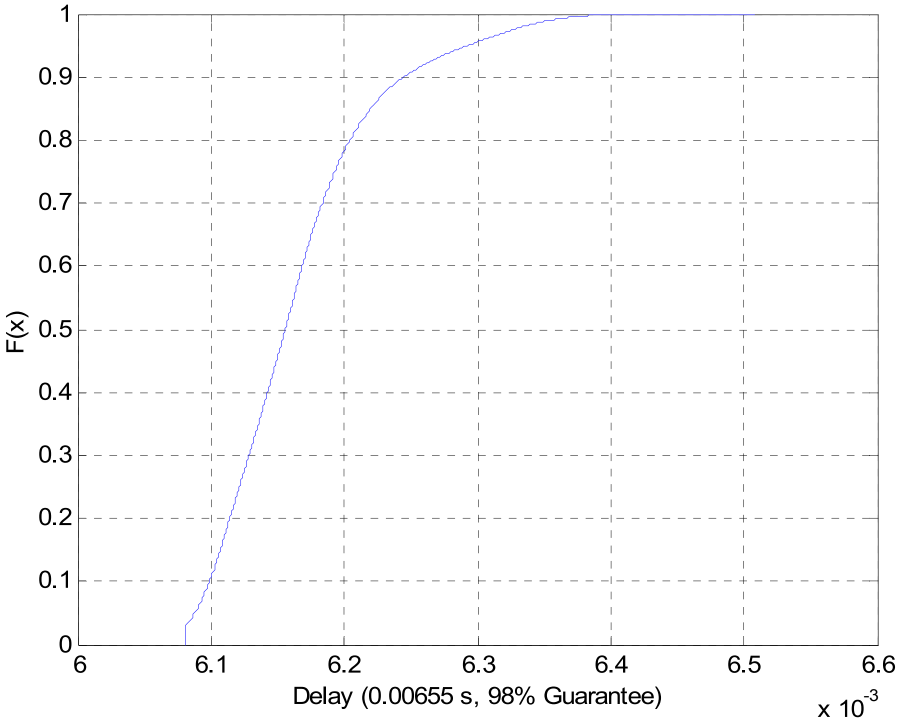


Figure 4.18: Voice Packet Delay Distribution (Edge 5 – Sink 9)

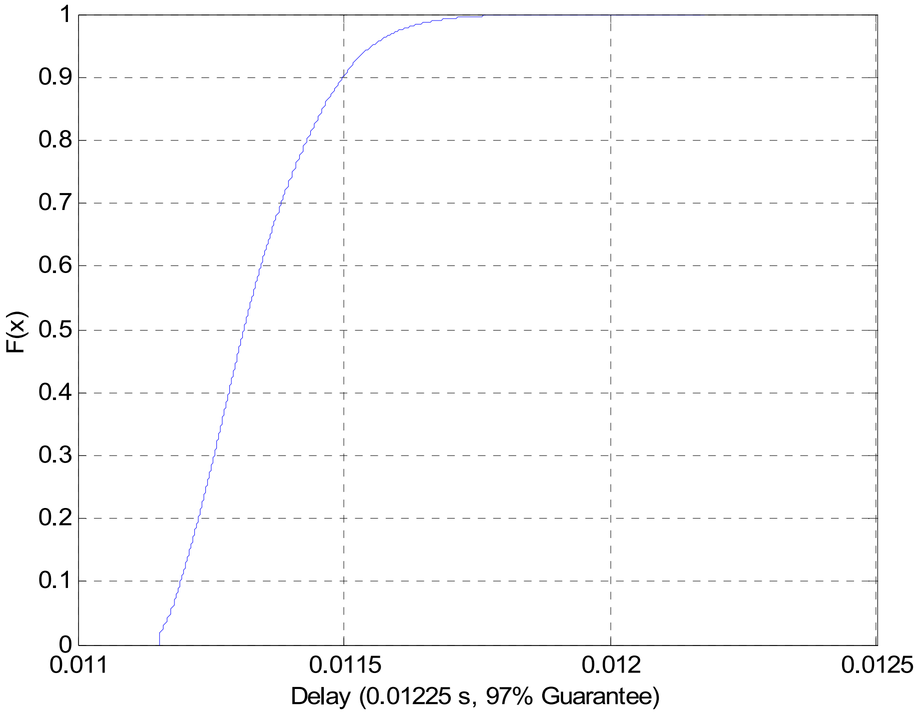


Figure 4.19: Video Packet Delay Distribution (Edge 1 – Sink 10)

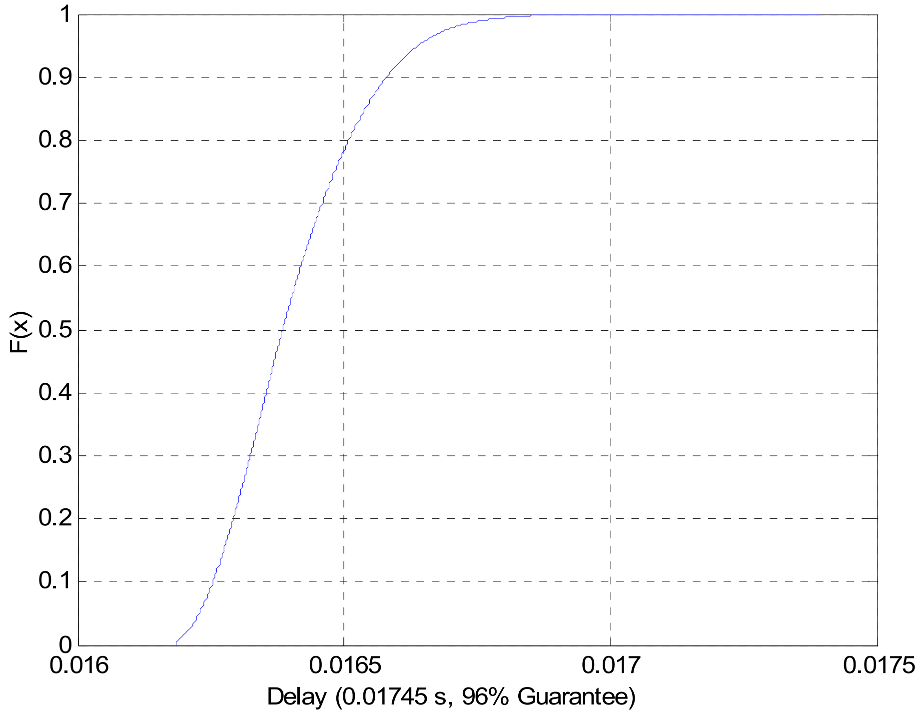


Figure 4.20: Video Packet Delay Distribution (Edge 1 – Sink 11)

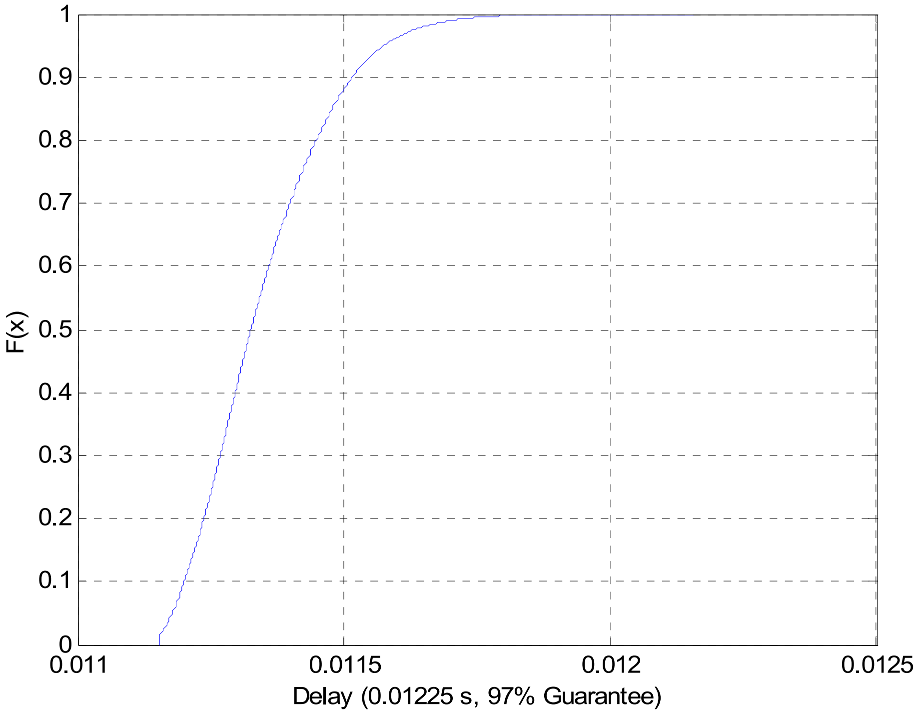


Figure 4.21: Video Packet Delay Distribution (Edge 2 – Sink 10)

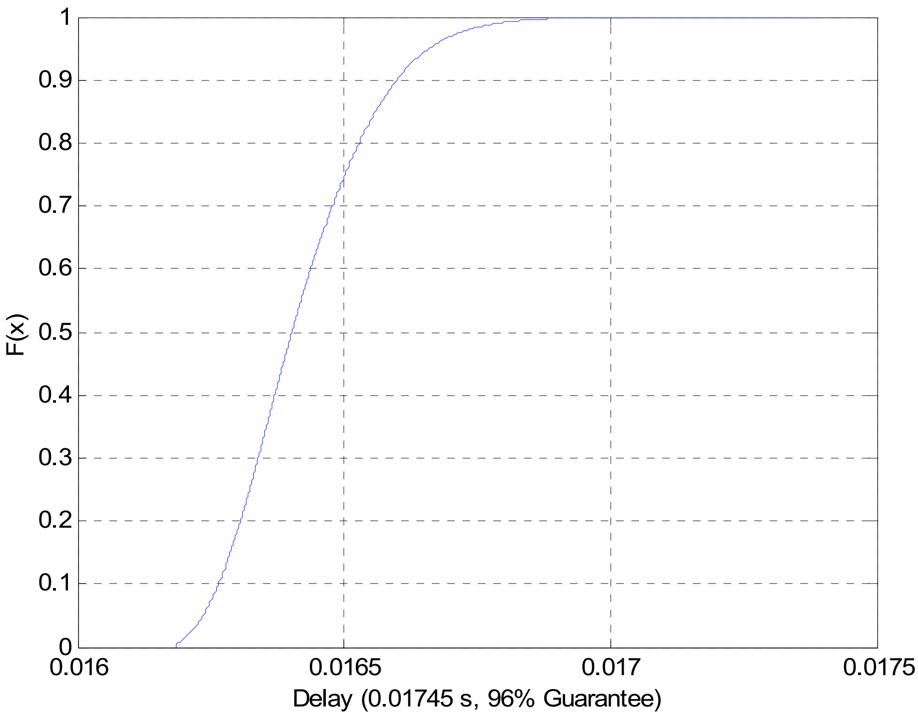


Figure 4.22: Video Packet Delay Distribution (Edge 2 – Sink 11)

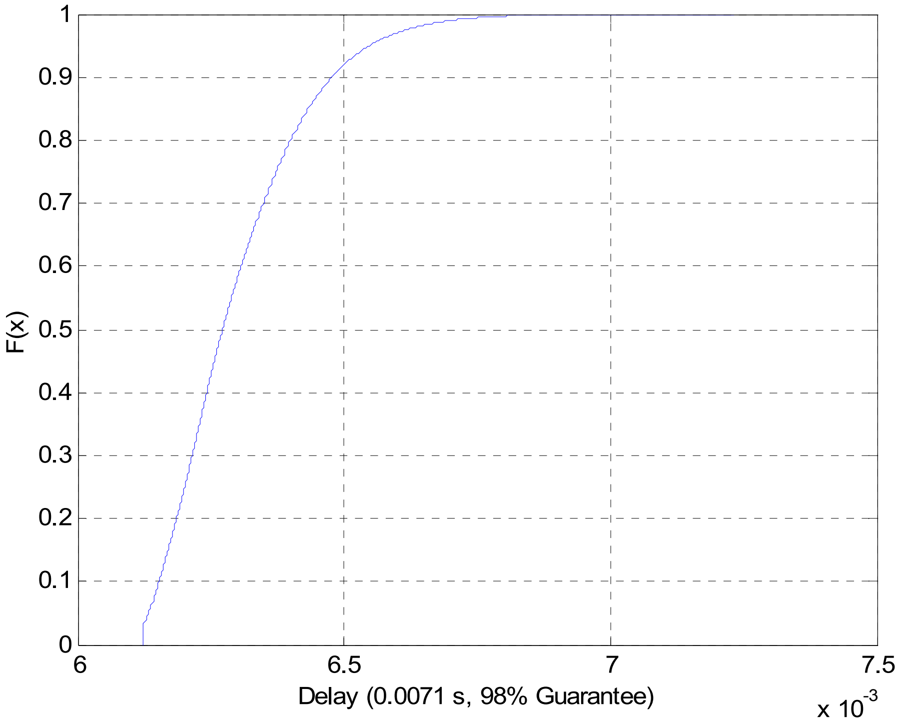


Figure 4.23: Video Packet Delay Distribution (Edge 3 – Sink 10)

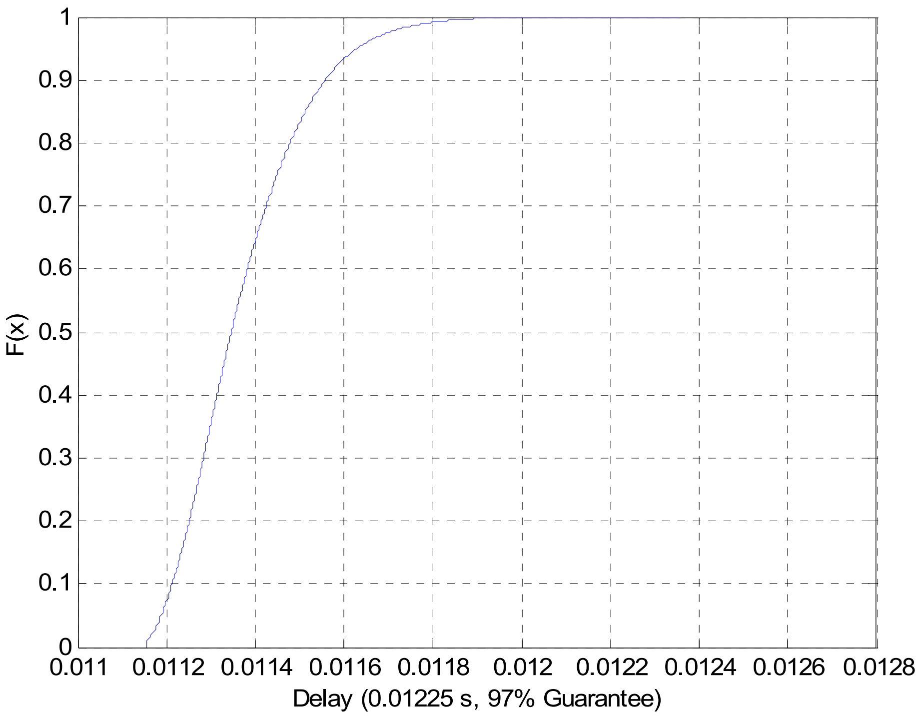


Figure 4.24: Video Packet Delay Distribution (Edge 3 – Sink 11)

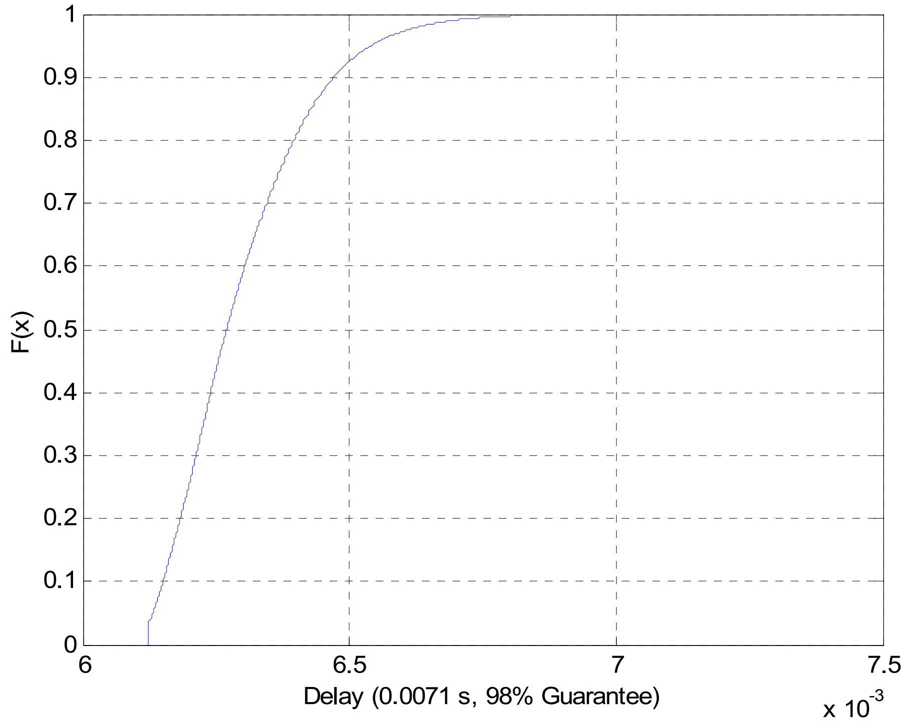


Figure 4.25: Video Packet Delay Distribution (Edge 4 – Sink 11)

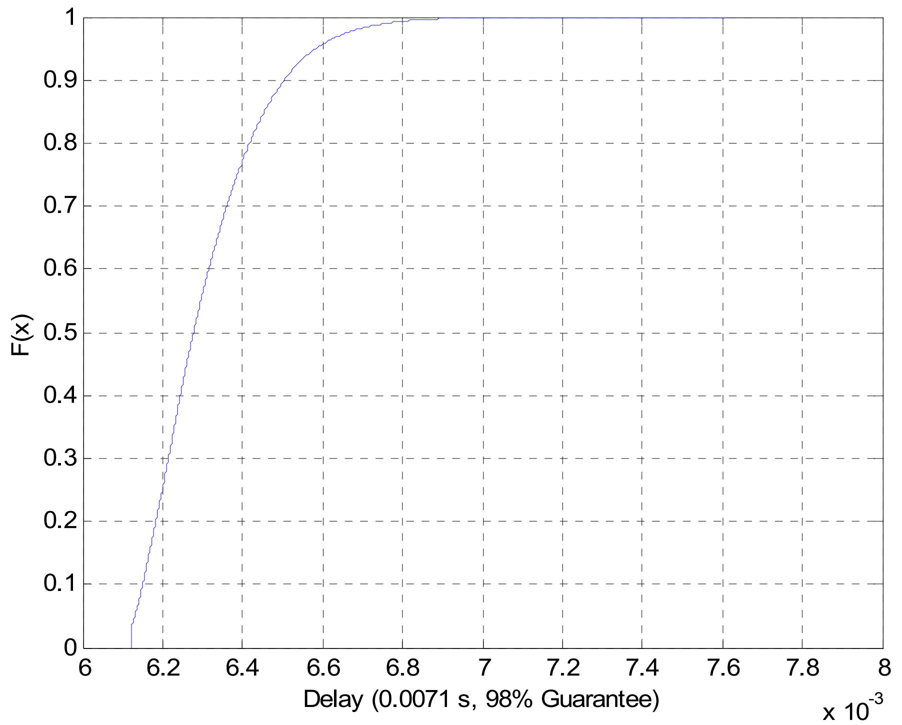


Figure 4.26: Video Packet Delay Distribution (Edge 5 – Sink 9)

From these figures, we see that the admission control scheme provides the statistical guarantee of the packet delay very well for different source-destination pairs, and the over-delayed packet percentage is below the bound set in the simulation, although it is a bit conservative. Actually, since we know the buffer size and the MTU (Maximum Transfer Unit) size, we can directly compute the worst-case end-to-end delay a voice packet may experience in the network. However, this is not applicable for video packets because it is not the highest priority class in the system.

Delay requirement and equivalent bandwidth allocation are two criteria in this admission control scheme. Each criteria can provide one admission region respectively, and the actual admission region should be the intersection of these two regions. Specifically, we should partition, either statically or dynamically, the loss probability and end-to-end delay violation probability among the routers traversed by each of the connections. One of the important functions involved in the admission control is the QoS routing, which is responsible for the task of finding a suitable path satisfying the two criteria from the source to the destination. If no such path is found, the connection request will be rejected.

In the simulation, all the admission control decisions are made by one bandwidth broker. However, in a real system, it will be a huge burden for a single node to handle all the works, so a distributed control system is needed, where several administrative entities cooperate together to manage the admission control in the whole network.

4.6 Conclusion

Most admission control schemes reviewed in Chapter 2 do not deal with the problem of providing both packet loss and delay guarantees or consider the multiclass services environment such as in a DiffServ network. In this chapter, we have studied and investigated our simple admission control scheme for multiclass traffic in the single-hop and multi-hop environment. Both resource allocation and measurement based methods are used in this scheme, so it is more like a hybrid one, but it is different from the hybrid scheme introduced in Chapter 2.

The simulation results show that the scheme proposed can provide effective loss guarantee as well as the statistical guarantee of end-to-end delay. The satisfying system utilization also addresses the efficiency of the scheme. For the high priority real-time traffic, i.e., voice and video, a large buffer will not impose much increase in the packet delay, but it can diminish the loss ratio significantly. On the contrary, the lower priority non-real-time traffic needs a relative larger buffer size to achieve a low loss ratio. The simple summation of four classes' equivalent bandwidth will result in system underutilization, while the reduced-service-rate approximation can push up the utilization at the cost of loss bound violation in some situations.

Chapter 5

End-to-End Admission Control

The end-to-end connection spans over the U_u wireless interface and the wireline part from Node B to the end user through the Core Network and the external DiffServ IP network. In this chapter we will investigate the problem of the admission control and the provision of QoS guarantees in end-to-end communications.

5.1 Admission Control in UMTS

As described in chapter 3, it is the UMTS Bearer Service that provides the UMTS QoS, which consists of two parts: Radio Access Bearer Service and Core Network Bearer Service. The Radio Access Bearer Service provides transport of signaling and user data between MT and CN I_u Edge Node with the negotiated QoS requirement, which is mostly based on the wireless interface. The Core Network Bearer Service of UMTS connects the CN I_u Edge Node with the CN Gateway to the external network, e.g., DiffServ networks. This is to control and utilize the backbone network to provide the negotiated UMTS bearer service and it should support different services for a variety of QoSs.

The Radio Access Bearer Service is realized by the Radio Bearer Service and the I_u Bearer Service. The role of the Radio Bearer Service is to cover all aspects of the radio interface transport, which uses UTRA FDD in our scheme, and the I_u Bearer Service provides the transport between UTRAN and CN. The Core Network Bearer Service uses the generic Backbone Network Service, which is not specified in UMTS but may use the existing standard, such as DiffServ.

As shown in Figure 3.1, the end-to-end connection covers the wireless interface from the mobile station to Node-B, wireline part from Node-B to GGSN in UMTS domain and the external DiffServ network. The admission control policy is implemented in each domain along the end-to-end path. Only if each domain accepts the request, then the new connection is admitted. The admission control in DiffServ network has been introduced in chapter 4, and here we will discuss the admission control policy in the UMTS domain.

5.1.1 WCDMA Wireless Interface Admission Control

The Radio Resource Management (RRM) is responsible for the utilization of the air interface resources, which is needed to provide the negotiated quality of service. It consists of handover, power control, packet scheduling, admission control, etc. Before admitting a new connection, the admission control entity will check whether enough radio resources (e.g., power, channel, and code) are available, and the admittance of a new connection should not degrade the QoS of existing connections below their QoS requirements. If the request is accepted, a radio access bearer in the radio access network will be established. The admission control functionality is implemented in RNC and it is executed separately for the uplink and downlink directions. The radio bearer service will be setup only if both the uplink and downlink admission control admit the connection. Otherwise the request is rejected.

In chapter 3, several admission control schemes in CDMA wireless interface have been presented. They are based on power, interference and load, etc. The uplink transmission is considered in our scheme. The outage probability and packet loss ratio are the primary criteria for admission control algorithm in the wireless domain. Gilhousen et al. [36] give the uplink capacity for a multiple cell CDMA system. Vannithamby and Sousa [37] analyze the capacity of a variable spreading gain CDMA system, where two traffic classes with different data rates, i.e., different spreading gains are considered.

There are four QoS traffic classes in UMTS, and each class has its own QoS requirement in the wireless interface. Considering the situation where each user has only one type of service and one connection, the following equations give the expressions for the SIR of each class [40].

$$
\begin{aligned}
&\frac{S_1 G_1}{p_1(N_1-1)S_1 + (p_{2l} M N_2 S_{2l} + p_{2h} N_2 S_{2h}) + p_3 N_3 S_3 + p_4 N_4 S_4 + E[I_{\text{intercell}}] + \eta} = \gamma_1 \ , \\
&\frac{S_{2l} G_{2l}}{p_1 N_1 S_1 + [p_{2l} M (N_2-1) S_{2l} + p_{2h} (N_2-1) S_{2h}] + p_3 N_3 S_3 + p_4 N_4 S_4 + E[I_{\text{intercell}}] + \eta} = \gamma_2 \ , \\
&\frac{S_{2h} G_{2h}}{p_1 N_1 S_1 + [p_{2l} M (N_2-1) S_{2l} + p_{2h} (N_2-1) S_{2h}] + p_3 N_3 S_3 + p_4 N_4 S_4 + E[I_{\text{intercell}}] + \eta} = \gamma_2 \ , \\
&\frac{S_3 G_3}{p_1 N_1 S_1 + (p_{2l} M N_2 S_{2l} + p_{2h} N_2 S_{2h}) + p_3 (N_3-1) S_3 + p_4 N_4 S_4 + E[I_{\text{intercell}}] + \eta} = \gamma_3 \ , \\
&\frac{S_4 G_4}{p_1 N_1 S_1 + (p_{2l} M N_2 S_{2l} + p_{2h} N_2 S_{2h}) + p_3 N_3 S_3 + p_4 (N_4-1) S_4 + E[I_{\text{intercell}}] + \eta} = \gamma_4 \ .
\end{aligned}
\tag{5.1}
$$

S_1, S_{2l}, S_{2h}, S_3 and S_4 are the received power of voice, video low-bit-rate, video high-bit-rate, interactive and background sources at the Node-B, respectively,

$G_1, G_{2l}, G_{2h}, G_3, G_4$, $\gamma_1, \gamma_{2l}, \gamma_{2h}, \gamma_3, \gamma_4$ and $p_1, p_{2l}, p_{2h}, p_3, p_4$ are the spreading gains, target SIRs, and activity factors of each class. N_1, N_2, N_3, N_4 are the number of users in the four classes in every cell, respectively, and M is the number of low-bit-rate exponential on-off sources in each video user, which is eight in our scheme. η is the thermal noise power and $E[I_{\text{intercell}}]$ is the expectation of the inter-cell interference.

In equation (5.1), the numerator is the product of the spreading gain and the received power at Node-B. The denominator is the total interference which includes four parts: (1) the intra-cell interference from its own class, (2) the intra-cell interference from other classes, (3) the inter-cell interference, and (4) the thermal noise. For voice, interactive and background services, each user occupies only one dedicated channel when it has data to transmit, while one video user has eight low-bit-rate sources and one high-bit-rate source, and each source uses its own channel to transmit the data, so multiple dedicated channels are needed. The desired received power of each traffic class at the Node-B can be obtained from equation (5.1) as follows.

$$S_{2l} = \frac{G_1/\gamma_1 + p_1}{G_{2l}/\gamma_2 + p_{2l}M + p_{2h}\dfrac{G_{2l}}{G_{2h}}} \cdot S_1 \ ,$$

$$S_{2h} = \frac{G_{2l}}{G_{2h}} \cdot S_{2l} \ , \tag{5.2}$$

$$S_i = \frac{G_1/\gamma_1 + p_1}{G_i/\gamma_i + p_i} \cdot S_1 \ , \qquad i \in \{3,4\} \ .$$

If we set the power value of the voice traffic assuming perfect power control, then the received power of the other three traffic classes can be computed from the above equations. Perfect power control means that the received power at Node-B of each user in the same class is at the same value.

In [38], an analytical formulation of the outage probability in terms of bit error rate for multiclass services in the uplink of a wideband CDMA cellular system is presented. This paper investigates the case where different traffic classes have different spreading gains. The outage probability analysis of variable bit rate multiclass services in the uplink of wideband WCDMA is presented in [39], where a variable bit rate (VBR) source is modeled by a continuous-time Markov chain with finite states. Multiple spreading codes with the same data rate are used by each VBR source. The video source in our admission control is modeled by a two-dimensional Markov chain which consists of high- and low-bit-rates exponential on-off sources, so the analytical formulation of the outage probability in [40] is the main approach to obtain the WCDMA admission region. Multiple low-bit-rate spreading codes with the same data rate and a high-bit-rate spreading code with a higher data rate are used by each video source.

In the uplink direction of the WCDMA wireless interface, the admission region is based on the outage probability performance in terms of bit error ratio specification and packet loss ratio for the multiclass services. We consider three schemes with different mobile user service provisioning and their corresponding admission regions.

In Scheme 1, each mobile user can use only one type of service and set up only one connection at the same time, i.e., single-connection per user. The received power of each class at Node-B is set to a fixed value according to equation (5.2) no matter how many number of users there are in the system. For all the four traffic classes, no packet retransmission is employed, which means that the outage packet is discarded and considered as packet loss. Each traffic class has its BER requirement with the corresponding SIR value. Of course, no packet retransmission for data traffic is not realistic. However, this serves as a starting point where this scheme is extended to schemes 2 and 3 which consider packet retransmission for data traffic. Through the method introduced in [40], we can calculate the outage probability in terms of each class's SIR requirement with a specified combination of the number of connections for each class. If the outage probability specifications for each class are set, the WCDMA wireless interface admission region can be obtained. To allow a continuous flow of the schemes' description, the relevant equations and the detail procedure are presented in Appendix A.

In Scheme 2, it is also the case of single-connection per user. However, the dynamic power control is applied, if the number of users changes, the necessary received power of each traffic class is recomputed. Furthermore, packet retransmission and transmission buffer are provided for the non-real-time traffic, i.e., Interactive and Background. Therefore packet loss is due to buffer overflow and transmission failure after reaching the maximum retransmission time. The method to obtain the wireless admission region of this scheme and Scheme 3 is described in [46].

In Scheme 3, each mobile user can use multiple types of services and set up more than one connection simultaneously, i.e., multi-connection per user. There are several groups, each with different service type combination. Each user is in one of the groups. Due to the orthogonality characteristics between the different channels from the same mobile station, the intra-cell interference consists of only the received powers from other mobile users. Dynamic power control is also employed. If the number of users changes, the received power of each traffic class is recomputed, and the same service type traffic in different groups may have different power values. Packet retransmission and transmission buffer are provided for the non-real-time traffic. The approach to obtain the wireless admission region is the same with Scheme 2. In fact, Scheme 2 is a subset of Scheme 3, i.e., four groups of user combination with each group has only one of the four UMTS service classes.

5.1.2 UMTS Wireline Network Admission Control

The wireline network of UMTS consists of three parts: I_{ub} interface between Node-B and RNC, I_u interface between UTRAN and CN, and the backbone of Core Network. In chapter 3, the QoS management functions for UMTS are presented, and here we give an brief introduction of the above wireline parts and relevant protocols.

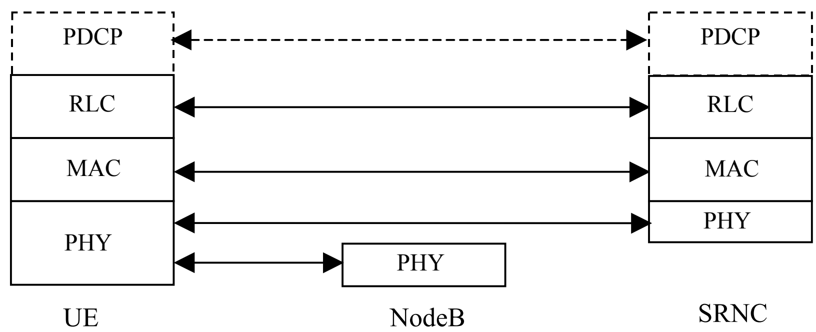


Figure 5.1: Protocol Termination for DCH, User Plane

Figure 5.1 shows the protocol termination for DCH for the user plane. The Serving RNC (SRNC) is the top-most macro-diversity combining and splitting function for the FDD mode [42]. The I_{ub} should provide the delivery of the Frame Protocol PDU between Node-B and RNC. When there are some data to be transmitted, DCH data frames are transferred in every transmission time interval from the SRNC to the Node-B for the downlink transfer, and from Node-B to the SRNC for the uplink transfer. The point-to-point connection between the Node-B and the UTRAN is considered as the Last Mile Link, which shall be modeled as an infinite server providing a fixed service rate [43].

The UMTS architecture showed in Figure 3.1 encompasses the packet-switched network through the General Packet Radio Service evolution. GPRS is the network architecture to provide efficient access to external packet data networks from the cellular networks, which introduces a backbone network based on IP. This backbone network consists of new networks nodes as well as traditional packet network nodes such as routers. Two main network elements, SGSN and GGSN, interact with each other and with the existing cellular network elements over a set of interfaces and implement a variety of functions. The SGSN, connected to the RNC in UTRAN over I_u interface, relays the data packet through the IP backbone to the GGSN on the other side of the core network, while the GGSN performs as the edge router providing connectivity to the external IP networks and handles the resource provisioning.

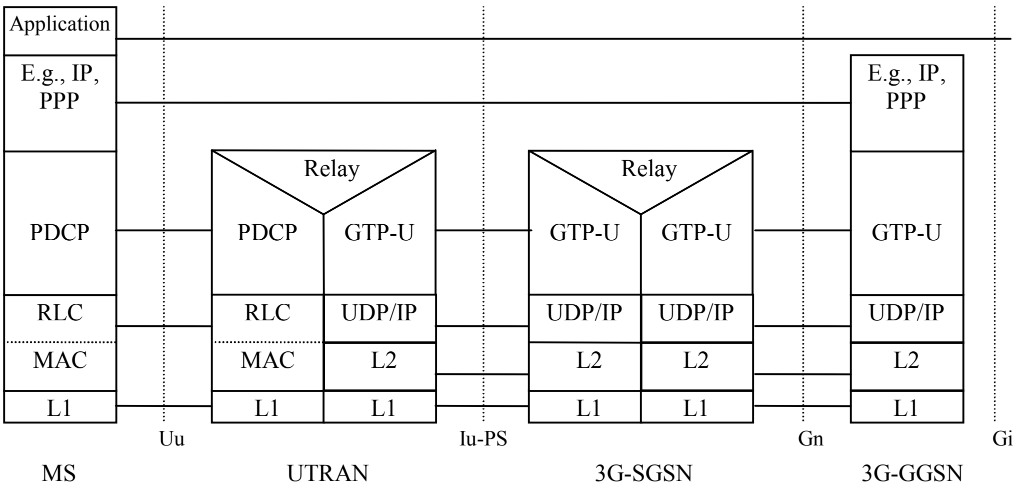


Figure 5.2: MS-GGSN User Plane with UTRAN

In order to send or receive packet data, the UE must activate a Packet Data Protocol (PDP) context, a virtual connection between the mobile station and the GGSN, and establish a packet data session and mobility management states in the MS as well as in the network. Figure 5.2 presents the MS-GGSN user plane with UTRAN [44], which includes the protocols for the data transmission between different network elements. A GPRS Tunneling Protocol (GTP) tunnel for the user plane (GTP-U) is defined for each PDP context in the GSNs and each RAB in the RNC, which tunnels user data between UTRAN and the SGSN, and between the GSNs in the backbone network. A GTP tunnel is necessary to forward packets between MS and the external packet data networks, and it encapsulates all PDP PDUs. The GTP-U protocol is implemented by SGSN and GGSN in the Backbone and by RNC in the UTRAN, other network elements does not need to be aware of the GTP.

The DiffServ IP network architecture is employed here as the transportation technique over the core network and I_u interface, so we can impose the admission control policy for the four QoS classes of traffic introduced in chapter 4. The difference is that the packet size is much larger due to the GTP, the UDP and IP overhead. The average rate, the burst size and the equivalent bandwidth of the traffic will increase as a consequence.

5.2 End-to-End QoS Architecture

To provide end-to-end QoS, it is necessary to manage the QoS within each domain. For an end-to-end IP connection, the QoS management functions in UMTS network comprises of three parts: IP BS Manager, Translation/Mapping function and Policy Decision Function (PDF) [45].

The IP BS Manager uses standard IP mechanisms to manage the IP bearer services, which may include the support of the DiffServ Edge function and the RSVP function. The Translation/Mapping function provides the inter-working between the mechanisms and parameters used within the UMTS bearer service and those used within the IP bearer service, and interacts with the IP BS Manager, in short, the QoS mapping between UMTS QoS and external IP QoS metrics. PDF is a logical policy decision element which uses standard IP mechanisms to implement Service Based Local Policy (SBLP) in the IP bearer layer, including policy-based admission control, etc.

The GGSN supports DiffServ edge functionality and is compliant to the IETF specification for Differentiated Services. Upon receiving an end-to-end connection request, the GGSN sends an bearer authorization request to the PDF. The PDF will authorize the request according to the stored SBLP and the network status information. If the connection request is accepted in the final decision, the PDF will authorize and enable the resource for the connection depending on its QoS requirement.

The QoS issues of DiffServ network have been introduced in chapters 2 and 4, which include PHB, QoS mapping, admission control, etc. In the DiffServ IP network, it is necessary to perform resource management to ensure the QoS guarantee and the management should be performed through an interaction with the UTMS network. Within the UMTS network, the resource management performed in the admission control decision is under the direct control of the UMTS system. If UMTS uses external IP network resources as part of its bearer service (e.g., backbone bearer service), it is also necessary to inter-work with that network.

The use of DiffServ in the UMTS backbone network implies that the QoS mapping may not be needed between the CN and the external DiffServ IP network. If the UMTS system employs the fully DiffServ IP-based architecture and the same QoS techniques as the outside world, the internal and external QoS bearer services can be merged into one entity.

5.3 End-to-End Admission Control Strategy

In this section, we will describe the end-to-end admission control policy which spans from the WCDMA wireless interface, through the UMTS core network, and ends up at the DiffServ IP network.

As introduced in the previous section, the individual admission control is implemented in each domain. Only if all the domains accept the connection request and the end-to-end QoS requirement can be guaranteed, then the connection can be admitted. Since we are investigating the WCDMA uplink, the first step to an end-to-end admission control algorithm is to check the admission region of the wireless interface when the new connection request arrives. If the wireless admission region test is successful, then it comes to the second stage, i.e., the admission control in the wireline networks of UMTS. As the DiffServ IP architecture is employed in the UMTS core network and I_u interface, the algorithm introduced in chapter 4 can be applied. This includes the equivalent bandwidth and delay bound examination. The final step is the admission control in the external DiffServ IP network, as described before.

If all the domains have admitted the connection, the end-to-end QoS metrics are examined, i.e., the end-to-end packet loss ratio and the statistical delay bound must be guaranteed. After all the above stages have been performed, the final admission control decision is made. Failure in any stage will cause the request to be rejected. Otherwise, it will be accepted and the necessary resources in each domain are allocated to the connection.

5.4 End-to-End Simulation

We will investigate an end-to-end scenario simulation in this section, where the admission control scheme described in the previous section is applied. The simulation topology of the UMTS system is shown in Figure 5.3. There are a total of nine square-sized WCDMA cells in the system. The length of each cell is 1000 m. The center one is the reference cell, which receives the inter-cell interference from the surrounding eight cells. Each cell has the same simulation parameters, e.g., user combination, distribution and arrival rate.

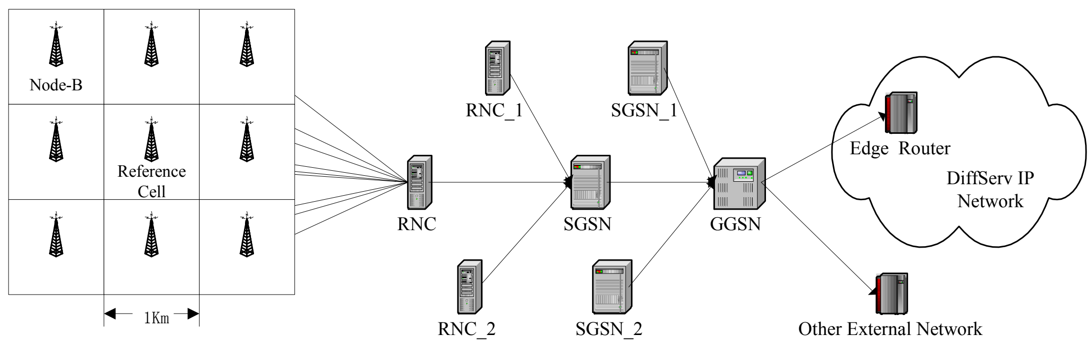


Figure 5.3: Simulation Topology of UMTS System

The packets received at the Node-Bs are sent to RNC, through SGSN, GGSN and finally to the external DiffServ IP network. The RNC_1, RNC_2, SGSN_1, SGSN_2 are the other UMTS network elements whose traffic are from the other cells or Radio Network Subsystem (RNS) and they act as the cross traffic in the UMTS system. Their packets are transmitted to the other external networks. The topology of the DiffServ IP Network in the figure is the same as that shown in Figure 4.10, and the GGSN is connected to Edge Router 1. According to the three different approaches introduced in section 5.1.1, we have three simulation scenarios with different wireless admission regions.

5.4.1 Single-Connection without Retransmission

The wireless admission region of scheme 1 is used in this simulation. Each cell has four types of mobile station users with uniform distribution. The users are Voice, Video, Interactive and Background, whose traffic models have been presented in Table 4.2 and Table 4.5. Each user receives only one type of service at a time and the transmission buffer size is only one packet. The system has no retransmission, which means that if one packet is transmitted during the outage state, it is discarded by Node-B. For the voice, interactive and background users, each connection occupies only one DCH channel, while each video user need one high-rate and eight low-rate DCH channels due to its traffic characteristics.

Table 5.1: Simulation Parameters of Wireless Interface

	Voice	Video	Interactive	Background
Convolution	1/2	1/2	1/2	1/2
Bit Error Ratio	10^{-2}	10^{-2}	10^{-3}	10^{-3}
SIR Target	2 dB	2 dB	3 dB	3 dB
Outage/Loss Threshold	10^{-2}	10^{-2}	10^{-3}	10^{-3}
Physical Channel Rate	60 kbps	60 kbps (high) 30 kbps (low)	120 kbps	240 kbps
Spreading Gain	64	64 (high) 128 (low)	32	16
Mean Inter-arrival Time	2.0 s	0.2 s	20.0 s	2.0 s
Codec and Packetization Delay	20 ms	30 ms	20 ms	20 ms

Propagation Model	Propagation Constant (μ)	4
	Lognormal Shadow Deviation (σ)	6 dB
Thermal Noise (η)	-103.2 dBm (4.8×10^{-14} W)	

The wireless interface parameters are shown in Table 5.1. Because perfect power control of the mobile stations is assumed, the received power at the Node-B for each user in the same class is of the same value. In the simulation, we set the received power of voice at Node-B 1 dB lower than the thermal noise power. Thus the powers of other traffic classes can be obtained using equation (5.2). Since there is no retransmission in the wireless interface, we consider the outage probability as the packet loss ratio. With above parameters and the equations in Appendix A, we can obtain an admission region, and the wireless interface admission control is based on this region.

Table 5.2: Simulation Parameters of Wireline Networks

<table>
<tr><th rowspan="2"></th><th rowspan="2">Link</th><th rowspan="2">Server Output Link Rate (bps) and Prop Delay</th><th colspan="4">Buffer (pkts)</th></tr>
<tr><th>Voice</th><th>Video</th><th>Interactive</th><th>Background</th></tr>
<tr><td rowspan="4">UMTS Wireline Network</td><td>Node-B –> RNC</td><td>2M (5 ms)</td><td>Infinite</td><td>Infinite</td><td>Infinite</td><td>Infinite</td></tr>
<tr><td>RNC –> SGSN</td><td>10M (2 ms)</td><td>10000</td><td>10000</td><td>10000</td><td>10000</td></tr>
<tr><td>SGSN –> GGSN</td><td>15M (2 ms)</td><td>5</td><td>8</td><td>10</td><td>60</td></tr>
<tr><td>GGSN –> External Network</td><td>20M (2 ms)</td><td>4</td><td>5</td><td>6</td><td>8</td></tr>
<tr><td rowspan="2">DiffServ IP Network</td><td>Edge Router</td><td>10M (1 ms)</td><td>5</td><td>8</td><td>15</td><td>60</td></tr>
<tr><td>Core Router</td><td>28M (5 ms)</td><td>4</td><td>5</td><td>8</td><td>10</td></tr>
<tr><td colspan="7"></td></tr>
<tr><td></td><td colspan="3">Reference Cell –> Sink 10</td><td colspan="3">Reference Cell –> Sink 11</td></tr>
<tr><td>Voice Delay (ms)</td><td colspan="3">63.7</td><td colspan="3">68.8</td></tr>
<tr><td>Video Delay (ms)</td><td colspan="3">115.5</td><td colspan="3">120.7</td></tr>
</table>

Table 5.2 shows the wireline network simulation parameters. The DCH frame received at Node-B is transmitted to the RNC through the Last Mile Link, modeled as an infinite server providing a fixed service rate [43]. The frames are combined and reassembled into packets at the RNC, attached with GTP, UDP, IP overhead (288 bits)

and sent to the SGSN. The frame loss at Node-B and RNC should be avoided, because it will result in a packet recovery problem. Therefore we set very large buffers at both nodes. After the packet is switched to the GGSN, its GTP, UDP, and IP overheads are removed and it is sent to the external DiffServ IP network.

The packet loss ratio target of each wireline network router is the same as that in chapter 4, i.e., 10^{-2} for real-time traffic (Voice and Video) and 10^{-3} for non-real-time traffic (Interactive and Background). Thus the end-to-end packet loss ratio for a connection is approximately the loss ratio in the wireline networks plus the outage probability in WCDMA radio interface. A 99 percent delay guarantee is set at each router, thus the end-to-end delay bound violation probability is approximately the sum of all the violation probabilities at each router along the connection path. The destination nodes for the packets from the reference cell consists of Sink 10 and Sink 11 in the DiffServ IP network, and the end-to-end delay bound settings are presented in Table 5.2.

Table 5.3: Wireless Interface Simulation Results (Scheme 1)

	Voice	Video High	Video Low	Interactive	Background
Outage Threshold	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-3}	1.0×10^{-3}
Packet Outage Probability	1.5×10^{-3}	1.6×10^{-3}	1.3×10^{-3}	8.9×10^{-4}	9.6×10^{-4}

The total end-to-end simulation time is 12500 seconds, the earlier stage is the warm-up period and the numerical data is recorded from 6500 seconds onward. If the wireless channel is in outage, i.e., the SIR measured at the Node-B is lower than its threshold, the BER of the packet received will exceeds its threshold and the error correction of the data may not be achieved. Therefore the packet is discarded and lost.

Table 5.3 shows the outage probability thresholds used to compute the admission region and the corresponding simulation results. From these results we can see that the outage probability of each class traffic is well-controlled and within its threshold. Actually the number of users in each class has great effect on the outage probabilities, especially in some critical state, where the outage probability rises dramatically due to even a small increase in the number of users.

Table 5.4: Wireline Network Packet Loss Ratio (Scheme 1)

	Voice		Video		Interactive		Background		Total
	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Util
SGSN	1.8×10^{-4}	0.29	9.2×10^{-4}	0.41	5.3×10^{-5}	0.08	3.0×10^{-4}	0.09	0.87
GGSN	7.6×10^{-4}	0.24	2.9×10^{-4}	0.36	2.4×10^{-4}	0.14	1.6×10^{-5}	0.09	0.86
Edge 1	1.1×10^{-6}	0.09	3.5×10^{-4}	0.52	4.0×10^{-6}	0.16	7.2×10^{-4}	0.1	0.86
Core 6	9.6×10^{-8}	0.04	3.8×10^{-4}	0.61	3.3×10^{-4}	0.21	9.1×10^{-4}	0.08	0.93
Core 7	2.1×10^{-6}	0.04	5.1×10^{-4}	0.61	1.3×10^{-4}	0.2	8.2×10^{-4}	0.08	0.93
Core 8	2.1×10^{-6}	0.03	1.8×10^{-3}	0.61	5.2×10^{-4}	0.2	4.7×10^{-4}	0.07	0.92

Table 5.4 presents the mobile users packet loss ratio and utilization of the wireline network routers that the end-to-end connection traverses. Since it is assumed that Node-B and RNC have no packet loss, they are not listed. At each router, there is cross traffic sharing the resources with the connections from the WCDMA cells. The data in the table shows that the packet loss ratio for each traffic class at the routers is within the threshold, both in UMTS and DiffServ IP network. The end-to-end packet loss ratio is the sum of the loss ratio at each router and the outage probability in the reference cell. Since every router along the connection's path can guarantee its loss ratio as well as that in the WCDMA cells, the end-to-end packet loss can be well-estimated and controlled.

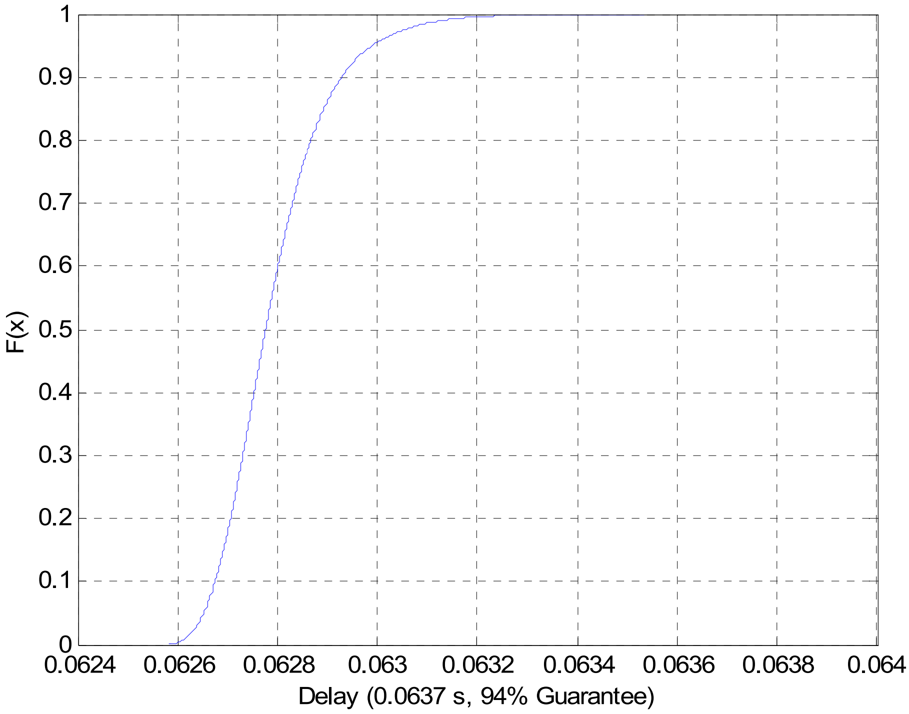


Figure 5.4: End-to-End Voice Packet Delay Distribution (Cell – Sink 10) (Scheme 1)

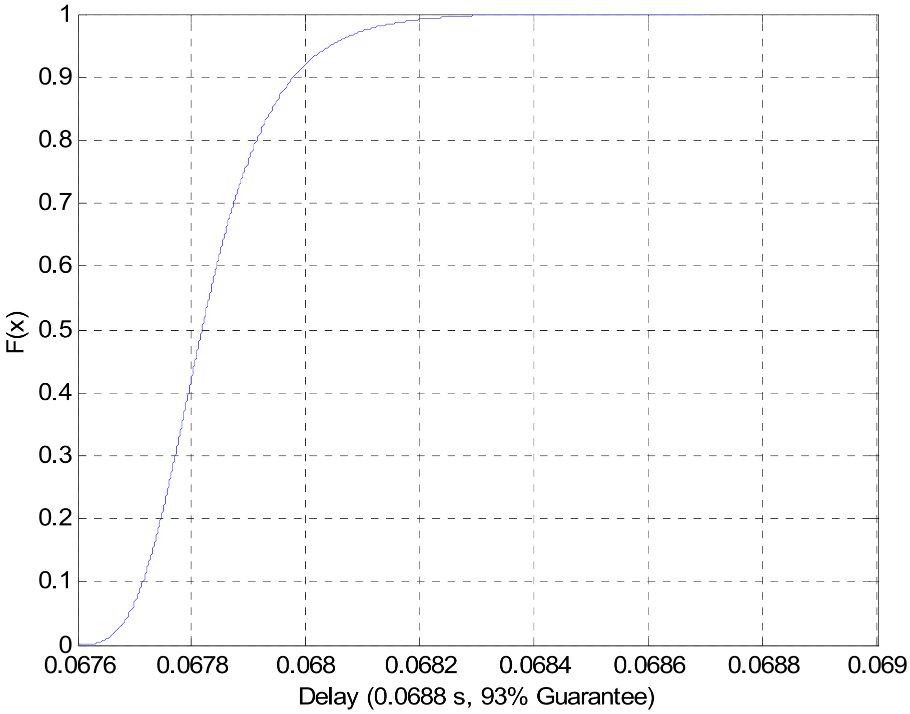


Figure 5.5: End-to-End Voice Packet Delay Distribution (Cell – Sink 11) (Scheme 1)

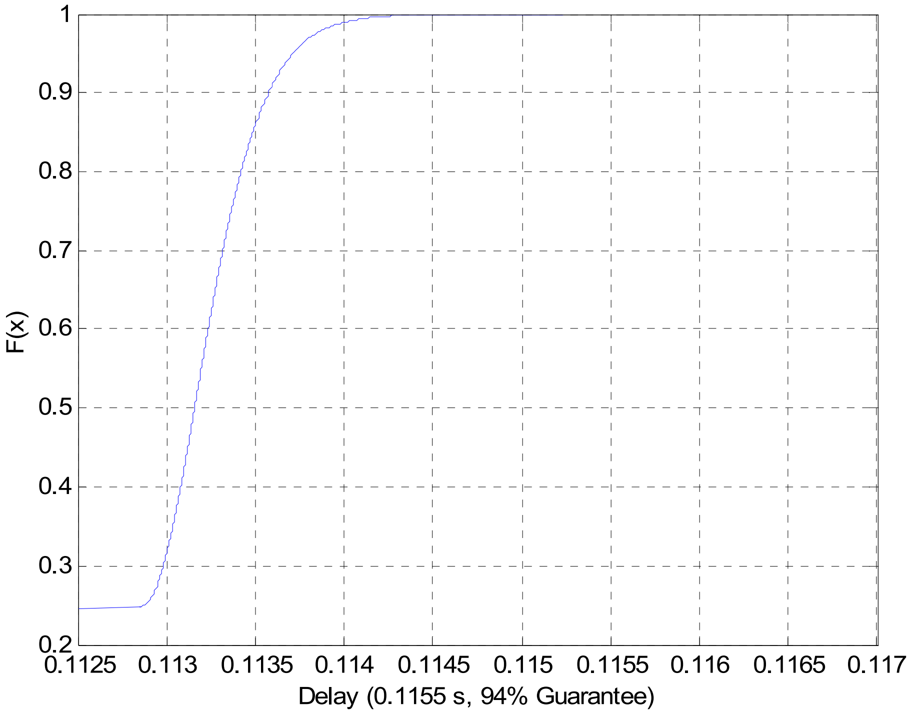


Figure 5.6: End-to-End Video Packet Delay Distribution (Cell – Sink 10) (Scheme 1)

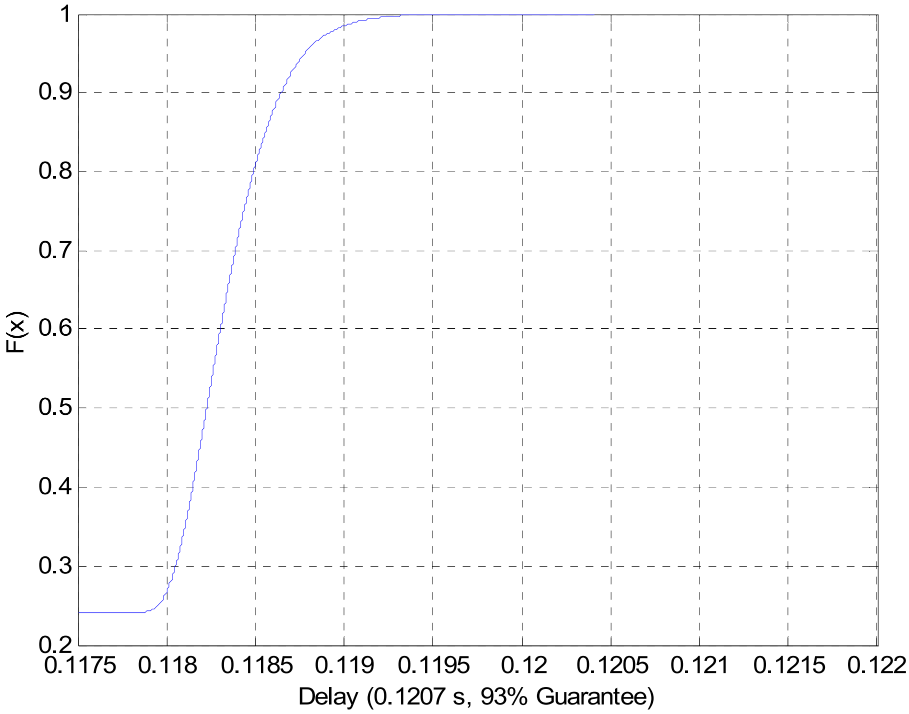


Figure 5.7: End-to-End Video Packet Delay Distribution (Cell – Sink 11) (Scheme 1)

Besides the packet loss ratio, the end-to-end packet delay also has a great effect on the real-time connection's QoS performance. Figures 5.4 to 5.7 present the end-to-end voice and video packet delay distribution, which are from the reference cell and received by the different destination nodes. In the simulation, the packetization delay at the mobile station, the transmission delay in the wireless channel and at each router are fixed values, so the delay jitter is caused by the queuing delay variation in the wireline network routers.

From the above figures, we can see that the admission control scheme can provide the end-to-end delay guarantee for connections from the WCDMA cell, as well as in the DiffServ network from the simulation results in chapter 4. In Figure 5.6 and 5.7, the minimum probability is not zero. This is because that the high- and low-data-rate sources for the video user use DCH channels of different rates to send its packets. The packet transmitted in the high-rate channel results in a lower end-to-end delay. Since we employ the strict priority scheduling in our scheme, the delay jitter of real-time traffic, which is caused by the buffer queuing at each router, only accounts for a very small part of the total end-to-end delay. The main delay is due to the transmission delay and other processing delays, and it is more significant in the wireless channel.

5.4.2 Single-Connection with Retransmission

The wireless admission region of scheme 2 is considered in this simulation. Each user can receive only one type of service and set up one connection at the same time. Dynamic power control is employed. When the number of users changes, the received power for each class is recalculated, and the minimum power values obtained are magnified by a specified rate to diminish the effect of thermal noise. This rate is set to 100 times in this simulation scenario.

The system in this scheme provides packet retransmission and transmission buffers for non-real-time traffic, where the Go-Back-N ARQ protocol is used. The sender receives an acknowledgement after a two-packet transmission period from the finish time of the packet transmission, and the number of maximum retransmission attempt is set at 3. The transmission buffer sizes for the Interactive and Background connections are 200 and 400 packets, respectively. In order to satisfy the outage and packet loss ratio requirements, the maximum numbers of voice, video, Interactive and Background users in each cell are 16, 4, 4 and 4, respectively. Other simulation parameters are the same as in the Tables 5.1 and 5.2.

Table 5.5: Wireless Interface Simulation Results (Scheme 2)

	Voice	Video High	Video Low	Interactive	Background
Outage (Packet Loss) Threshold	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-3}	1.0×10^{-3}
Outage (Packet Loss) Probability	1.2×10^{-3}	1.1×10^{-3}	6.8×10^{-4}	1.9×10^{-4}	9.9×10^{-5}

Table 5.5 presents the simulation results for the outage probability for Real-time Traffic and the packet loss ratio for Non-real-time Traffic. We can observe that the outage and packet loss targets are both achieved. For non-real-time traffic, i.e., Interactive and Background, the packet loss consists of two parts: (1) the transmission failure due to the outage in wireless channel after the maximum retransmission times, and (2) the transmission buffer overflow due to continuous retransmission. If the maximum retransmission number is too small, many packets will be discarded without enough attempts. On the contrary, if it is too large, the continuous retransmission may cause the user transmission buffer to overflow quickly, which results in a great packet

loss in a very short period. Furthermore, the large number of packet retransmissions of the non-real-time traffic will increase its on period length, then the outage probability in the wireless channel becomes larger and causes even more retransmissions, which is a vicious circle. In general, if the traffic load is light, more retransmission attempts can be enforced to avoid unnecessary packet discard, while in the situation where the wireless channel is in heavy load, smaller maximum retransmission times should be used to minimize the probability of buffer overflow and the channel QoS degeneration.

Table 5.6: Wireline Network Packet Loss Ratio (Scheme 2)

	Voice		Video		Interactive		Background		Total
	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Util
SGSN	1.9×10^{-4}	0.31	1.2×10^{-3}	0.41	2.5×10^{-5}	0.07	2.9×10^{-6}	0.06	0.84
GGSN	6.9×10^{-4}	0.25	2.5×10^{-4}	0.35	2.1×10^{-5}	0.14	3.2×10^{-6}	0.08	0.82
Edge 1	5.6×10^{-6}	0.13	6.6×10^{-5}	0.45	2.3×10^{-6}	0.14	2.5×10^{-5}	0.06	0.78
Core 6	3.7×10^{-5}	0.1	3.3×10^{-4}	0.54	1.6×10^{-5}	0.16	6.5×10^{-5}	0.11	0.91
Core 7	1.3×10^{-4}	0.1	5.5×10^{-4}	0.54	1.1×10^{-5}	0.15	3.1×10^{-5}	0.1	0.9
Core 8	1.4×10^{-4}	0.09	2.0×10^{-3}	0.55	1.2×10^{-4}	0.16	7.3×10^{-5}	0.11	0.9

The above table presents the packet loss ratio and utilization of the wireline network routers that the end-to-end connections traverse. The numerical results show that the packet loss ratio for each traffic class at routers is within its threshold, both in the UMTS and in the DiffServ IP network, which is the same conclusion as the results in Scheme 1. Figures 5.8 to 5.11 present the end-to-end voice and video packet delay distribution in this simulation for scheme 2. From the following figures, we can see that the admission control scheme can provide the end-to-end delay guarantee for connections from the WCDMA cell, as in the simulation results for scheme 1.

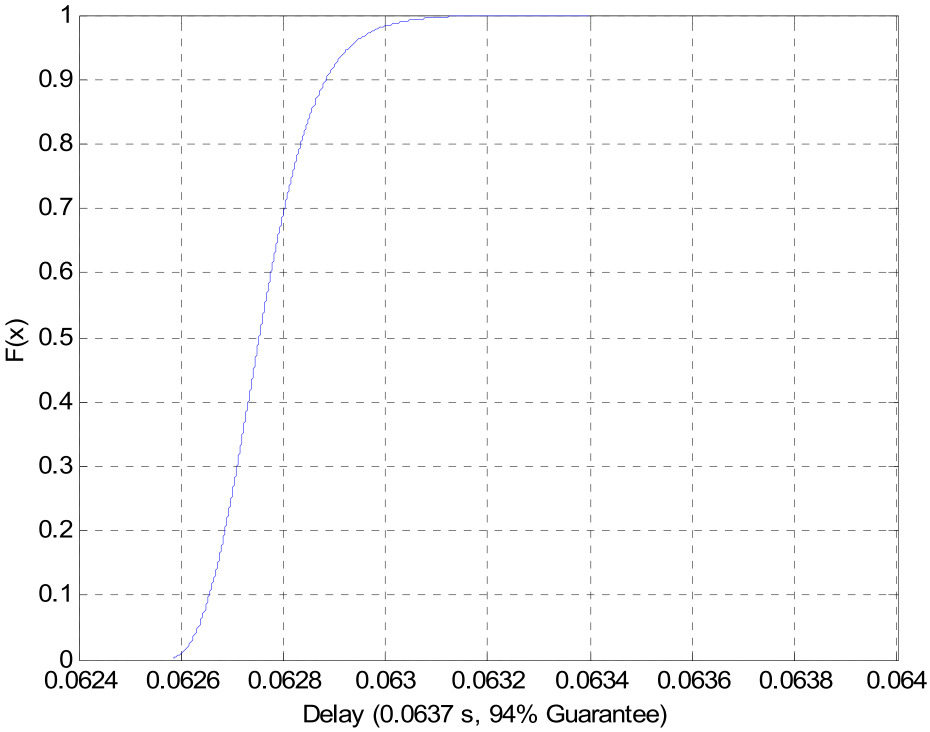


Figure 5.8: End-to-End Voice Packet Delay Distribution (Cell – Sink 10) (Scheme 2)

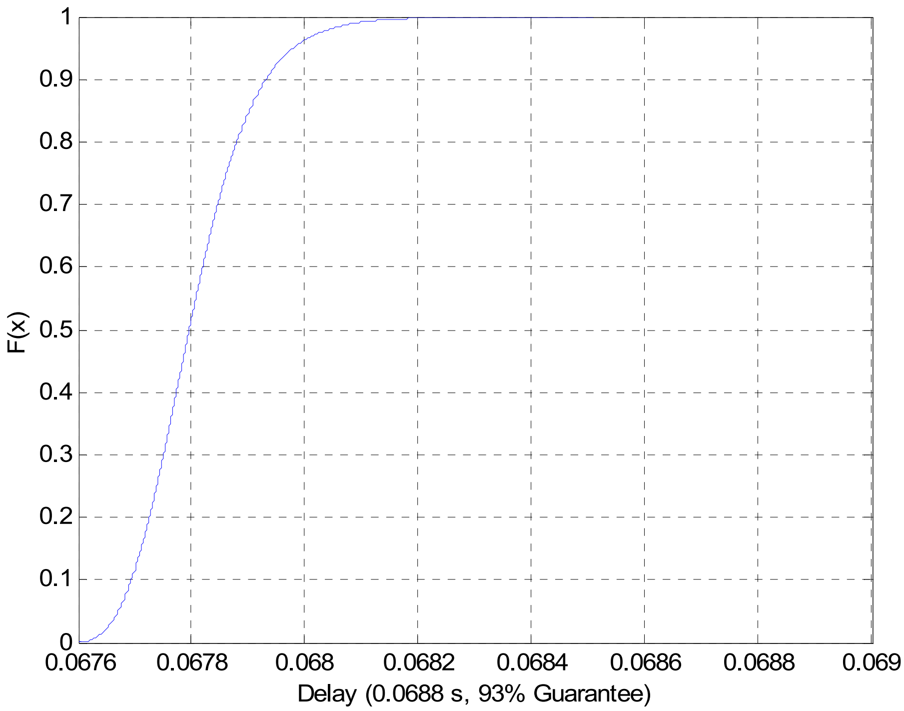


Figure 5.9: End-to-End Voice Packet Delay Distribution (Cell – Sink 11) (Scheme 2)

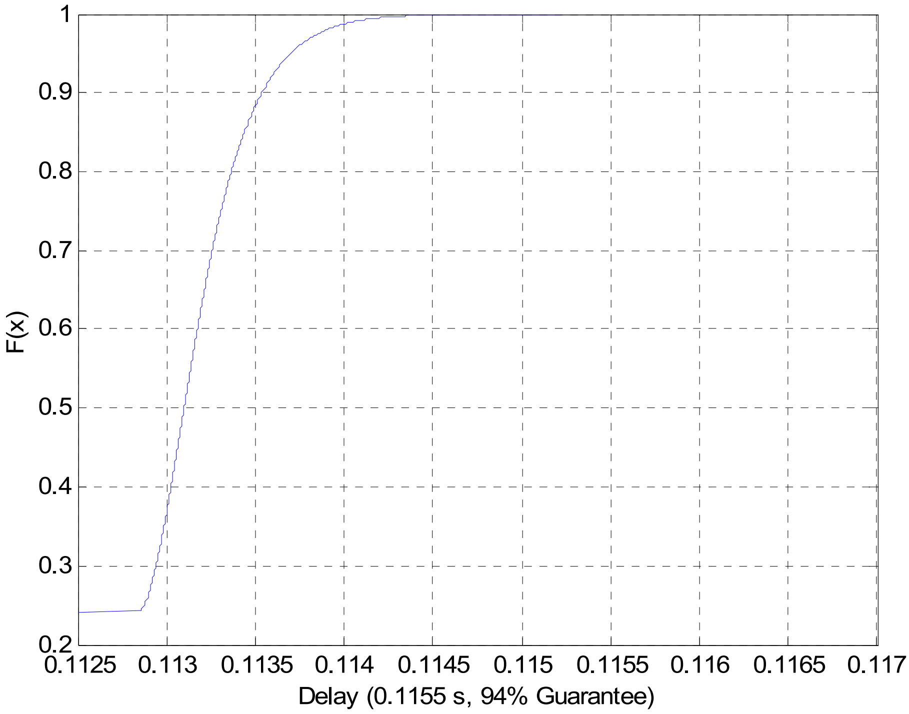


Figure 5.10: End-to-End Video Packet Delay Distribution (Cell – Sink 10) (Scheme2)

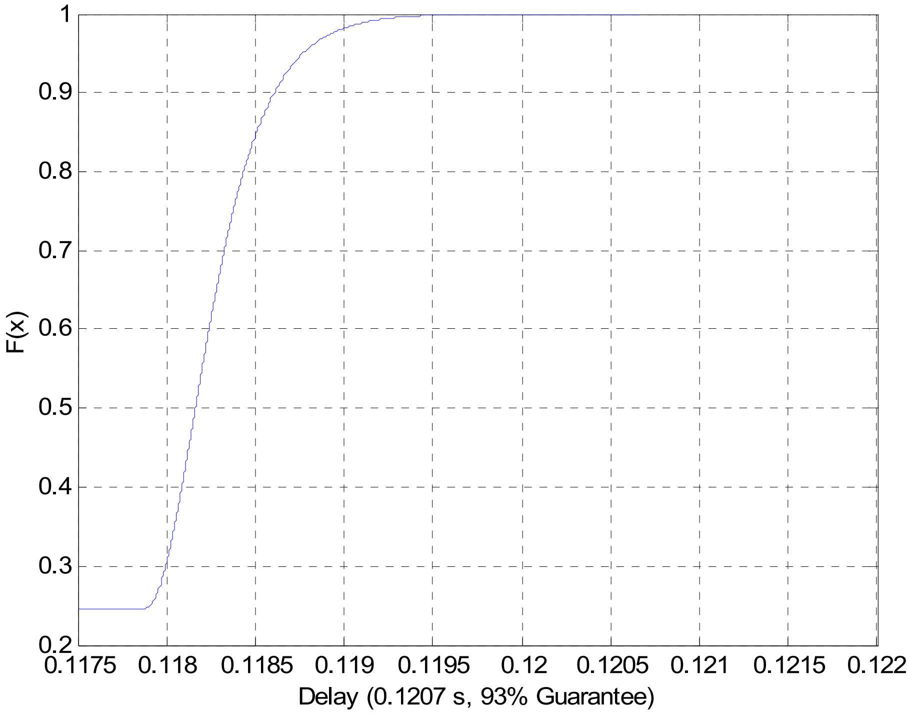


Figure 5.11: End-to-End Video Packet Delay Distribution (Cell – Sink 11) (Scheme 2)

5.4.3 Multi-Connection with Retransmission

The wireless admission region of scheme 3 is considered in this scenario. Each user can receive only one type of service combination at one time. The types and distribution of multi-code multi-service groups can be obtained through the network operator's statistics. In our simulation, there are four types of service combination groups, and they are shown in Table 5.7. Dynamic power control and non-real-time packet retransmission settings are the same as the case of the single-connection with retransmission. The following table gives the parameters in this simulation which are different from those of the single-connection case. Other parameters remain identical.

Table 5.7: Simulation Parameters in Multi-Connection with Retransmission

	Group 1	Group 2	Group 3	Group 4
Service Combination	1 Voice	1 Video	1 Voice + 1 Video	1 Interactive + 1 Background
Connection Holding Time	Exponential, $\mu = 300s$	Exponential, $\mu = 300s$	Exponential, $\mu = 300s$	Lognormal, Median = 300, Shape = 2.5
Mean Interarrival Time	1.0s	0.15s	0.1s	10.0s
	Interactive		Background	
Transmission Buffer (pkts)	200		400	

Table 5.8 presents the simulation results of the WCDMA wireless domain, which show that the outage probability and packet loss ratio for all the traffic classes are well-bounded. Compared to the scenario of single-connection without retransmission, the packet retransmission provides better packet loss guarantee for the non-real-time traffic in the wireless interface with the same channel quality.

Furthermore, the cancellation of the interference from the connections of the same mobile station can enhance the system capacity.

Table 5.8: Wireless Interface Simulation Results (Scheme 3)

	Voice	Video High	Video Low	Interactive	Background
Outage (Packet Loss) Threshold	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-2}	1.0×10^{-3}	1.0×10^{-3}
Outage (Packet Loss) Probability	6.6×10^{-3}	5.5×10^{-3}	4.2×10^{-3}	7.4×10^{-4}	8.3×10^{-4}

Table 5.9 presents the simulation results of the wireline network routers that the end-to-end connections traverse. It shows that the packet loss ratio for each traffic class at the routers is within the threshold, both in the UMTS and the DiffServ IP network, which is the same conclusions as with the results in the other scenarios in sections 5.4.1 and 5.4.2.

Table 5.9: Wireline Networks Simulation Results (Scheme 3)

	Voice		Video		Interactive		Background		Total
	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Loss Ratio	Util	Util
SGSN	2.6×10^{-4}	0.33	9.0×10^{-4}	0.37	3.8×10^{-5}	0.07	8.0×10^{-6}	0.06	0.82
GGSN	6.6×10^{-4}	0.25	2.0×10^{-4}	0.33	2.4×10^{-4}	0.14	3.0×10^{-6}	0.07	0.8
Edge 1	6.8×10^{-6}	0.15	5.2×10^{-5}	0.42	2.8×10^{-6}	0.13	6.4×10^{-6}	0.06	0.78
Core 6	6.0×10^{-5}	0.1	2.6×10^{-4}	0.52	9.2×10^{-6}	0.16	3.8×10^{-5}	0.11	0.9
Core 7	1.9×10^{-4}	0.1	4.5×10^{-4}	0.53	5.6×10^{-6}	0.15	1.2×10^{-5}	0.11	0.89
Core 8	2.3×10^{-4}	0.1	1.7×10^{-3}	0.52	9.5×10^{-5}	0.16	4.0×10^{-5}	0.12	0.9

Figures 5.12 to 5.15 show the end-to-end voice and video packet delay distribution in the multi-connection environment. From these figures, we can see that the admission control scheme can provide the end-to-end delay guarantee of connections from the WCDMA cell to the DiffServ network receivers.

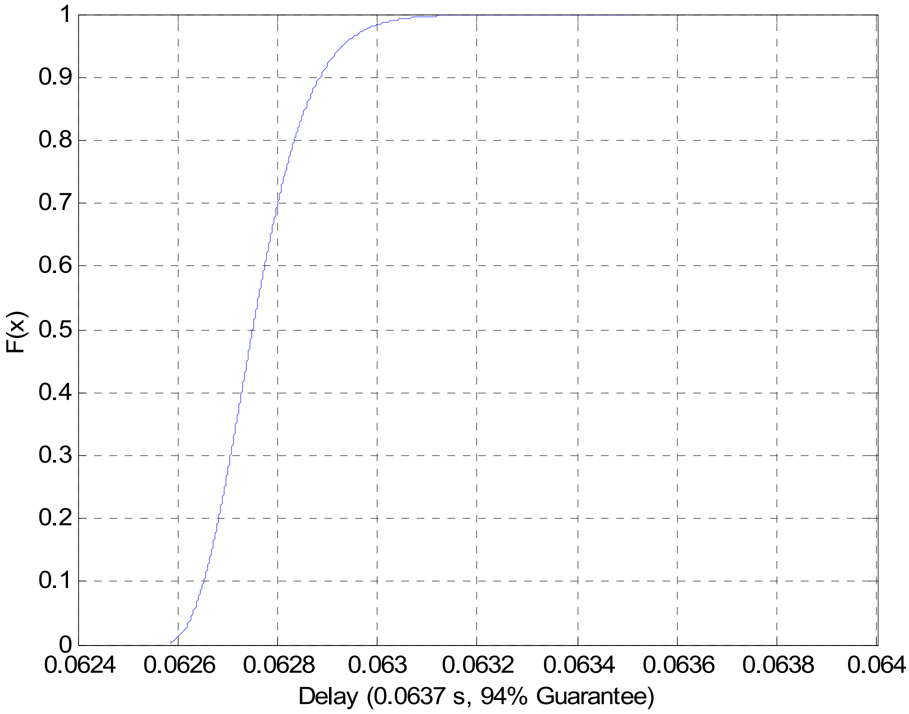


Figure 5.12: End-to-End Voice Packet Delay Distribution (Cell – Sink 10) (Scheme 3)

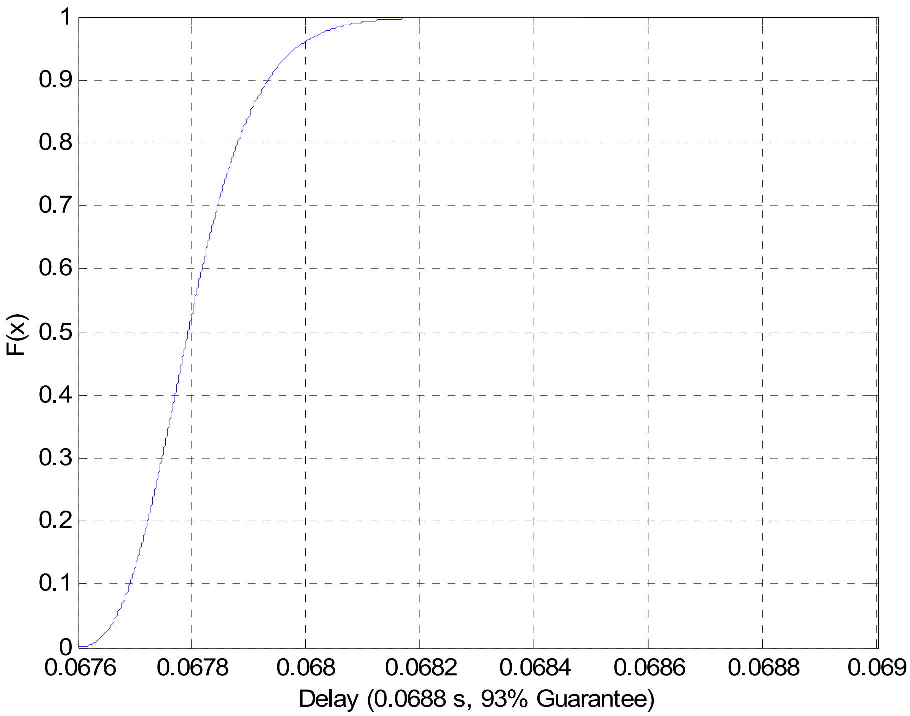


Figure 5.13: End-to-End Voice Packet Delay Distribution (Cell – Sink 11) (Scheme 3)

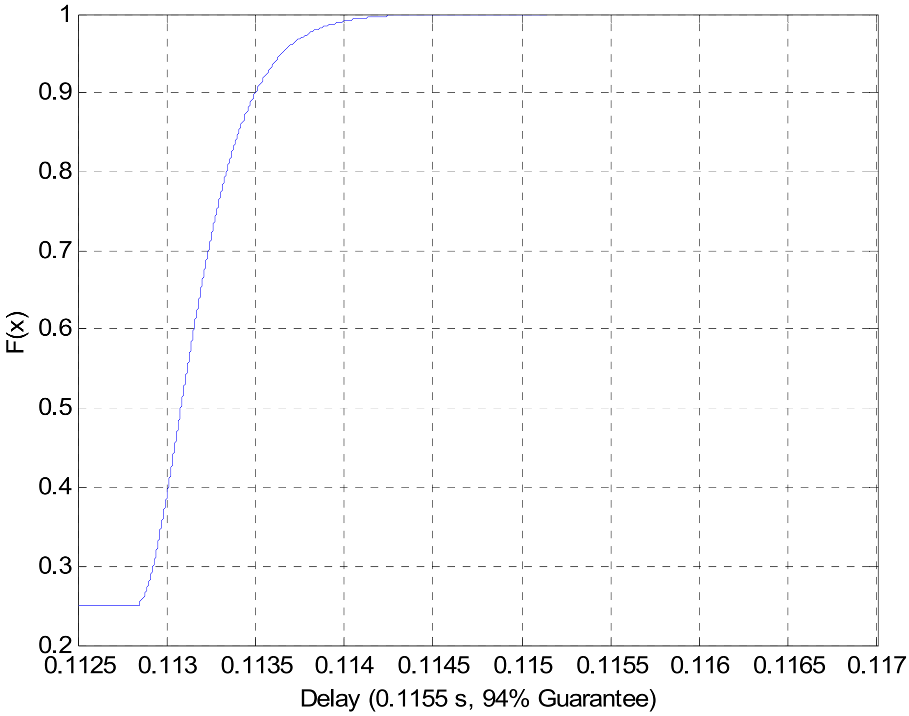


Figure 5.14: End-to-End Video Packet Delay Distribution (Cell – Sink 10) (Scheme 3)

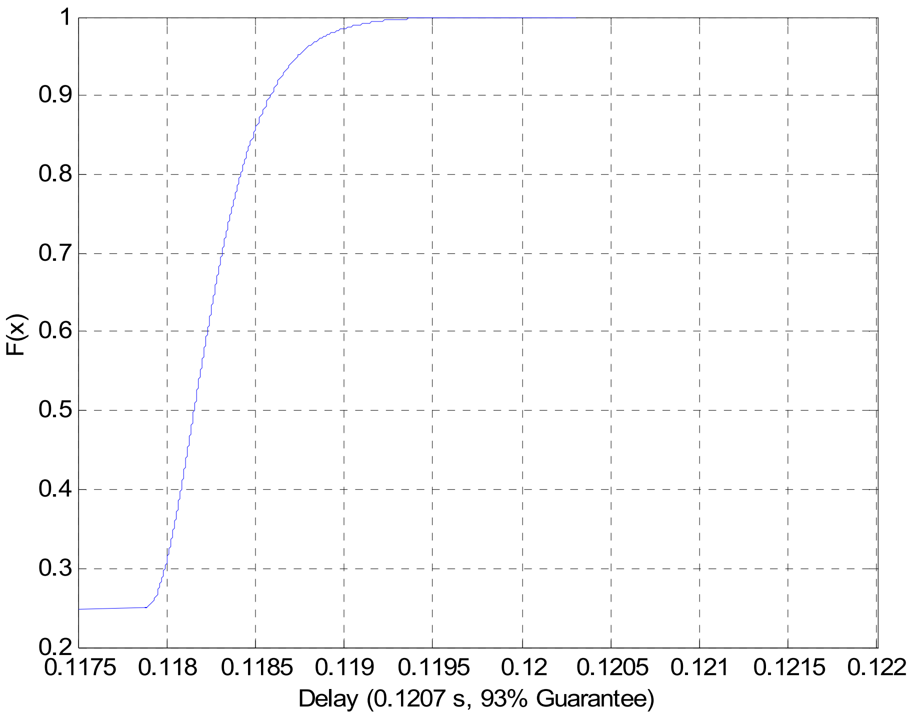


Figure 5.15: End-to-End Video Packet Delay Distribution (Cell – Sink 11) (Scheme 3)

5.5 Admission Control in Downlink Direction

The mobile service connections include uplink and downlink communications, so the admission control of the two directions should be implemented. The connection will be set up only if both the uplink and downlink admission control admit it. The downlink transmission is different from the uplink transmission in several aspects. The downlink transmission from Node-B is synchronous so that the intra-cell interference is only due to the multipath in the wireless channel. Because there is only one set of orthogonal codes used to support all the users, the downlink transmission also has a limit in terms of code assignments. The admission control in the downlink direction is very similar to the uplink scheme in this thesis and only differs in the wireless admission region, which can be referred to in [50]. In the downlink admission control, the WCDMA wireless admission region should be computed and the wireline domain of the scheme is the same as that of the uplink admission control.

5.6 End-to-End Admission Control Implementation

Figure 5.16 summarizes the flow chart of the end-to-end admission control scheme between the WCDMA wireless network and DiffServ IP wireline network.

When a new connection arrives, the wireless admission region, including both the uplink and the downlink directions, will be examined. If these tests are successful, then the DiffServ wireline network admission control in the two directions are enforced, which consists of the equivalent bandwidth admission region test and the delay bound estimation. After the above stages, the end-to-end QoS guarantees (e.g., packet loss ratio and delay) are checked. The connection will be admitted if the QoS guarantees are satisfied. The WCDMA uplink, downlink wireless admission region and the DiffServ wireline admission region (based on equivalent bandwidth without considering delay) can be computed beforehand and stored in a database. However, the overall wireline delay bound estimations for real-time traffic are not stored since they depend on the real-time measurements, unlike the fixed transmission delay in the wireless interface. End-to-end system designer can obtain information from the results in this chapter and the following scheme flow chart.

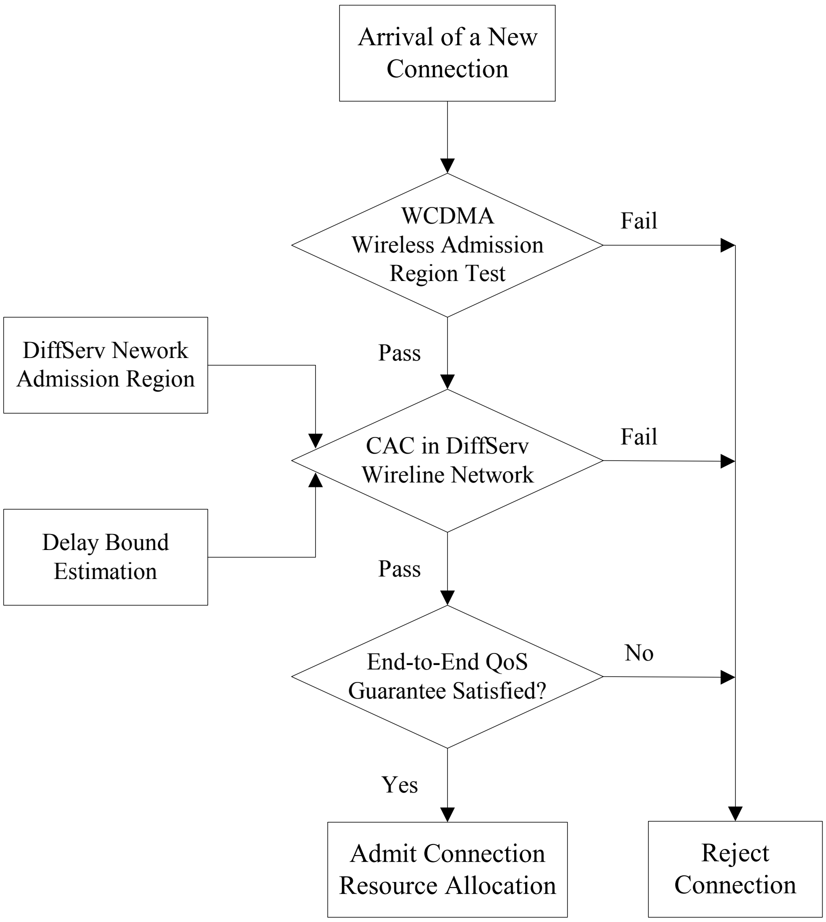


Figure 5.16: End-to-End Admission Control Scheme Flow Chart

5.7 Conclusion

The end-to-end admission control scheme from the WCDMA cell to the DiffServ IP network is investigated in this chapter. In the wireless admission region, we consider three different schemes, i.e., single-connection without retransmission, single-connection with retransmissions and multi-connection with retransmissions. Three different admission regions in the WCDMA cell are calculated based on these three schemes, respectively.

The admission control algorithm presented in chapter 4 is applied here in the UMTS and the DiffServ IP wireline networks. If both the wireless and wireline domains admit a connection request, and the end-to-end packet loss ratio and statistical delay can be guaranteed, the connection is admitted and the necessary resources are allocated in each domain according to its QoS requirements. The results of all the simulations show that the end-to-end admission control can provide the packet loss ratio and delay QoS guarantees of the connections from the wireless to the wireline networks, while the multi-connection scheme can achieve the largest system capacity and better packet loss guarantees for non-real-time traffic.

Chapter 6

Conclusion

6.1 Thesis Contribution

This thesis investigates the QoS provisioning issues of multiclass traffic traversing across a WCDMA mobile network and a wireline DiffServ IP network. It focuses on end-to-end admission control. Our main objective is to propose effective and efficient admission control algorithms for end-to-end delivery of multimedia information between mobile users and fixed network users with specified QoS guarantees.

The end-to-end connection spans over both a WCDMA wireless cellular network in a 3G system and a DiffServ IP based wireline network. Resource management and QoS guarantees in such a hybrid network is a challenging work due to the limited radio resources, poor wireless link quality, heterogeneous multimedia traffic and the different QoS mechanisms in each type of network.

As UMTS and DiffServ networks have different QoS classifications, we first define the QoS classes mapping between the two frameworks according to their QoS requirements, i.e., the Conversational, Streaming, Interactive and Background classes in the UMTS to the Expedited Forwarding and Assured Forwarding PHBs in the

DiffServ network.

A measurement-dependent resource allocation admission control scheme in a DiffServ network is proposed in the thesis and it attempts to inter-work with the four QoS classes of UMTS. In order to provide both packet loss ratio and statistical delay guarantee efficiently for the connections in the network, equivalent bandwidth and statistical delay bound estimation are employed. Through the management of equivalent bandwidth allocation, the packet loss ratio at each router can be bounded, with the delay bound estimation. The statistical delay control at the router can also be ensured such that the end-to-end packet loss ratio and statistical delay guarantee are satisfied.

From simulation results, we have studied the effect of buffer size on the four traffic classes with different priorities in the system. Under our simulation scenarios, the higher priority the traffic, the smaller the buffer size that is needed to provide the loss ratio guarantee. The necessary buffer size for the lower priority class, especially the Background traffic with the lowest priority, will increase quickly if the system load is close to full utilization. Moreover, the increase in buffer size of the real-time traffic results in a small delay increase due to their high priorities in the system. However, it can reduce the loss ratio significantly. Under the condition that there is a large number of connections, the Gaussian approximation application in the equivalent bandwidth management is less conservative and provide satisfactory system utilization.

We further apply the above scheme in end-to-end admission control, both in the UMTS DiffServ capable wireline network and the external IP network. The admission control in a WCDMA cell is based on the admission region calculated with the outage probability and packet loss ratio of each class in the wireless channel. Three different schemes, i.e., single-connection without retransmission, single-connection with retransmissions and multi-connection with retransmissions, are investigated. The wireless and wireline networks interact with each other in the admission control. If both domains have enough resources to support the new connection and the end-to-end QoS requirements can be guaranteed, the connection is admitted. The simulation results show that the schemes work well in end-to-end QoS guarantees under different traffic scenarios.

6.2 Future Work

This research work has led us to propose multiclass admission control algorithms in a DiffServ IP network and the interconnection of wireless (WCDMA) and wireline (DiffServ IP) networks with end-to-end QoS guarantees. However, there are still many research directions that can be studied in the future. The following outlines some of these directions which can extend the current work:

1. Scheduling algorithms have great influence on the QoS metrics, such as packet loss ratio and delay. In our scheme, a strict priority scheduling is employed. In future work, different scheduling algorithms can be used and their impact on QoS performance may be investigated.
2. In the QoS mapping between UMTS and DiffServ IP networks, we define only one type of EF PHB, i.e., the voice traffic. Actually we can further consider the service differentiation in the EF, which means multiple expedited forwarding PHBs can be defined in the DiffServ architecture, e.g., voice and video traffic.
3. Finally, the interworkings between the wireless and wireline networks on QoS management should be further investigated to improve the QoS performance of end-to-end connections.

Appendix

WCDMA Wireless Admission Region

We present an algorithm to calculate the WCDMA wireless admission region based on the outage probability of each traffic class and it is based on the analytical formulation of the outage probability with a source modeled by a two-dimensional Markov chain [40].

Equation (1) gives the expressions for the SIRs of each class, namely, Voice, Video, Interactive and Background traffic.

$$\frac{S_1 G_1}{p_1(N_1-1)S_1+(p_{2l}MN_2S_{2l}+p_{2h}N_2S_{2h})+p_3N_3S_3+p_4N_4S_4+E[I_{\mathrm{intercell}}]+\eta}=\gamma_1$$

$$\frac{S_{2l} G_{2l}}{p_1N_1S_1+[p_{2l}M(N_2-1)S_{2l}+p_{2h}(N_2-1)S_{2h}]+p_3N_3S_3+p_4N_4S_4+E[I_{\mathrm{intercell}}]+\eta}=\gamma_2$$

$$\frac{S_{2h} G_{2h}}{p_1N_1S_1+[p_{2l}M(N_2-1)S_{2l}+p_{2h}(N_2-1)S_{2h}]+p_3N_3S_3+p_4N_4S_4+E[I_{\mathrm{intercell}}]+\eta}=\gamma_2 \qquad (1)$$

$$\frac{S_3 G_3}{p_1N_1S_1+(p_{2l}MN_2S_{2l}+p_{2h}N_2S_{2h})+p_3(N_3-1)S_3+p_4N_4S_4+E[I_{\mathrm{intercell}}]+\eta}=\gamma_3$$

$$\frac{S_4 G_4}{p_1N_1S_1+(p_{2l}MN_2S_{2l}+p_{2h}N_2S_{2h})+p_3N_3S_3+p_4(N_4-1)S_4+E[I_{\mathrm{intercell}}]+\eta}=\gamma_4$$

Equation (2) gives the received power of each traffic class at Node-B.

$$S_{2l}=\frac{G_1/\gamma_1+p_1}{G_{2l}/\gamma_2+p_{2l}M+p_{2h}\dfrac{G_{2l}}{G_{2h}}}\cdot S_1$$

$$S_{2h}=\frac{G_{2l}}{G_{2h}}\cdot S_{2l} \qquad (2)$$

$$S_i=\frac{G_1/\gamma_1+p_1}{G_i/\gamma_i+p_i}\cdot S_1\,, \qquad i\in\{3,4\}$$

It is assumed that all the mobile users are uniformly distributed in the cell, so the expectation of the inter-cell interference $E[I_{\text{intercell}}]$ is expressed as:

$$
\begin{aligned}
E\left[\mathrm{I}_{\text{intercell}}\right] &= E\left[I_1 + (I_{2l} + I_{2h}) + I_3 + I_4\right] \\
&= S_1 p_1 \rho_1 \iint f\left(\frac{r_m}{r_d}\right) dA + S_{2l} M p_{2l} \rho_2 \iint f\left(\frac{r_m}{r_d}\right) dA + S_{2h} p_{2h} \rho_2 \iint f\left(\frac{r_m}{r_d}\right) dA \\
&+ S_3 p_3 \rho_3 \iint f\left(\frac{r_m}{r_d}\right) dA + S_4 p_4 \rho_4 \iint f\left(\frac{r_m}{r_d}\right) dA \\
&= (S_1 p_1 \rho_1 + S_{2l} M p_{2l} \rho_2 + S_{2h} p_{2h} \rho_2 + S_3 p_3 \rho_3 + S_4 p_4 \rho_4) \cdot \iint f\left(\frac{r_m}{r_d}\right) dA \\
&= (S_1 p_1 N_1 + S_{2l} M p_{2l} N_2 + S_{2h} p_{2h} N_2 + S_3 p_3 N_3 + S_4 p_4 N_4) \cdot \iint f\left(\frac{r_m}{r_d}\right) dA \cdot \frac{1}{A_{cell}} \quad .
\end{aligned} \tag{3}
$$

where ρ_i is the density of the i^{th} class user in the cell, A_{cell} is the area of the cell,

$$
f\left(\frac{r_m}{r_d}\right) = \left(\frac{r_m}{r_d}\right)^4 e^{\left(\frac{\sigma \ln 10}{10}\right)^2} \cdot Q\left[\frac{\ln 10}{10}\sqrt{2\sigma^2} + \frac{40\log(\frac{r_m}{r_d})}{\sqrt{2\sigma^2}}\right] \quad , \tag{4}
$$

σ is the standard deviation of the lognormal shadowing, r_d is the distance between inter-cell mobile user that is causing interference and the reference Node-B, and r_m is the distance between the inter-cell mobile user and its own Node-B. The variance of the inter-cell interference $Var[I_{\text{intercell}}]$ is shown as follows:

$$
\begin{aligned}
Var[I_{\text{intercell}}] &= Var[I_1 + (I_{2l} + I_{2h}) + I_3 + I_4] \\
&= S_1^{\,2} \rho_1 \int\int [p_1 g(\frac{r_m}{r_d}) - p_1^{\,2} f^2(\frac{r_m}{r_d})] dA \\
&+ S_{2l}^{\,2} \rho_2 \int\int [M p_{2l} [1 + (M-1) p_{2l}] g(\frac{r_m}{r_d}) - (M p_{2l})^2 f^2(\frac{r_m}{r_d})] dA \\
&+ S_{2h}^{\,2} \rho_2 \int\int [p_{2h} g(\frac{r_m}{r_d}) - p_{2h}^{\,2} f^2(\frac{r_m}{r_d})] dA \\
&+ S_3^{\,2} \rho_3 \int\int [p_3 g(\frac{r_m}{r_d}) - p_3^{\,2} f^2(\frac{r_m}{r_d})] dA \\
&+ S_4^{\,2} \rho_4 \int\int [p_4 g(\frac{r_m}{r_d}) - p_4^{\,2} f^2(\frac{r_m}{r_d})] dA
\end{aligned}
$$

$$= \begin{bmatrix} \left(S_1{}^2 N_1 p_1 + S_{2l}{}^2 N_2 M p_{2l}[1+(M-1)p_{2l}] + S_{2h}{}^2 N_2 p_{2h} + S_3{}^2 N_3 p_3 + S_4{}^2 N_4 p_4\right) \iint g(\frac{r_m}{r_d}) dA \\ -\left(S_1{}^2 N_1 p_1{}^2 + S_{2l}{}^2 N_2 (M p_{2l})^2 + S_{2h}{}^2 N_2 p_{2h}{}^2 + S_3{}^2 N_3 p_3{}^2 + S_4{}^2 N_4 p_4{}^2\right) \iint f^2(\frac{r_m}{r_d}) dA \end{bmatrix}$$

$$\cdot \frac{1}{A_{cell}} \quad , \tag{5}$$

where $Q(y) = \frac{1}{\sqrt{2\pi}} \int_y^\infty e^{-x^2/2} dx$ and

$$g\left(\frac{r_m}{r_d}\right) = \left(\frac{r_m}{r_d}\right)^8 e^{\left(\frac{\sigma \ln 10}{5}\right)^2} \cdot Q\left[\frac{\ln 10}{5}\sqrt{2\sigma^2} + \frac{40\log(\frac{r_m}{r_d})}{\sqrt{2\sigma^2}}\right] \quad . \tag{6}$$

The value of $\iint f\left(\frac{r_m}{r_d}\right) dA \cdot \frac{1}{A_{cell}}$, $\iint g(\frac{r_m}{r_d}) dA \cdot \frac{1}{A_{cell}}$ and $\iint f^2(\frac{r_m}{r_d}) dA \cdot \frac{1}{A_{cell}}$ can be obtained by numerical method or simulation and they should be a constant value only related to the cell shape.

After we get the expectation and variance of the inter-cell interference, the outage probabilities of the four traffic classes are presented as follow:

1. Voice

$$\begin{aligned} &\Pr(BER_1 \geq BER_1^*) = \Pr(SIR_1 \leq SIR_1^*) \\ &= \sum_{n_1=0}^{N_1-1} \sum_{n_{2l}=0}^{MN_2} \sum_{n_{2h}=0}^{N_2} \sum_{n_3=0}^{N_3} \sum_{n_4=0}^{N_4} Q\left(\frac{\delta_1 - \mu_1}{\sigma_1}\right) \times \binom{N_1 - 1}{n_1} p_1^{n_1} (1-p_1)^{N_1 - 1 - n_1} \times \binom{MN_2}{n_{2l}} p_{2l}^{n_{2l}} (1-p_{2l})^{MN_2 - n_{2l}} \times \\ &\binom{N_2}{n_{2h}} p_{2h}^{n_{2h}} (1-p_{2h})^{N_2 - n_{2h}} \times \binom{N_3}{n_3} p_3^{n_3} (1-p_3)^{N_3 - n_3} \times \binom{N_4}{n_4} p_4^{n_4} (1-p_4)^{N_4 - n_4} \quad , \end{aligned}$$

where

$$\begin{aligned} &\delta_1 = G_1/\gamma_1 - \eta/S_1 \quad , \\ &\mu_1 = n_1 + (n_{2l}S_{2l} + n_{2h}S_{2h} + n_3 S_3 + n_4 S_4)/S_1 + E[I_{\text{intercell}}]/S_1 \quad , \\ &\sigma_1 = \frac{\sqrt{Var[I_{\text{intercell}}]}}{S_1} \quad . \end{aligned} \tag{7}$$

2. Video (Low Bit Rate Channel)

$$\Pr(BER_{2l} \geq BER_{2l}^{*}) = \Pr(SIR_{2l} \leq SIR_{2l}^{*})$$
$$= \sum_{n_1=0}^{N_1} \sum_{n_{2l}=0}^{M(N_2-1)} \sum_{n_{2h}=0}^{N_2-1} \sum_{n_3=0}^{N_3} \sum_{n_4=0}^{N_4} Q\left(\frac{\delta_{2l}-\mu_{2l}}{\sigma_{2l}}\right) \times \binom{N_1}{n_1} p_1^{n_1} (1-p_1)^{N_1-n_1} \times \binom{M(N_2-1)}{n_{2l}} p_{2l}^{n_{2l}} (1-p_{2l})^{M(N_2-1)-n_{2l}} \times$$
$$\binom{N_2-1}{n_{2h}} p_{2h}^{n_{2h}} (1-p_{2h})^{N_2-1-n_{2h}} \times \binom{N_3}{n_3} p_3^{n_3} (1-p_3)^{N_3-n_3} \times \binom{N_4}{n_4} p_4^{n_4} (1-p_4)^{N_4-n_4} \quad ,$$

where

$$\delta_{2l} = G_{2l}/\gamma_2 - \eta/S_{2l} \quad ,$$
$$\mu_{2l} = n_{2l} + (n_1 S_1 + n_{2h} S_{2h} + n_3 S_3 + n_4 S_4)/S_{2l} + E[I_{\text{intercell}}]/S_{2l} \quad , \qquad (8)$$
$$\sigma_{2l} = \frac{\sqrt{Var[I_{\text{intercell}}]}}{S_{2l}} \quad .$$

3. Video (High Bit Rate Channel)

$$\Pr(BER_{2h} \geq BER_{2h}^{*}) = \Pr(SIR_{2h} \leq SIR_{2h}^{*})$$
$$= \sum_{n_1=0}^{N_1} \sum_{n_{2l}=0}^{M(N_2-1)} \sum_{n_{2h}=0}^{N_2-1} \sum_{n_3=0}^{N_3} \sum_{n_4=0}^{N_4} Q\left(\frac{\delta_{2h}-\mu_{2h}}{\sigma_{2h}}\right) \times \binom{N_1}{n_1} p_1^{n_1} (1-p_1)^{N_1-n_1} \times \binom{M(N_2-1)}{n_{2l}} p_{2l}^{n_{2l}} (1-p_{2l})^{M(N_2-1)-n_{2l}} \times$$
$$\binom{N_2-1}{n_{2h}} p_{2h}^{n_{2h}} (1-p_{2h})^{N_2-1-n_{2h}} \times \binom{N_3}{n_3} p_3^{n_3} (1-p_3)^{N_3-n_3} \times \binom{N_4}{n_4} p_4^{n_4} (1-p_4)^{N_4-n_4} \quad ,$$

where

$$\delta_{2h} = G_{2h}/\gamma_2 - \eta/S_{2h} \quad ,$$
$$\mu_{2h} = n_{2h} + (n_1 S_1 + n_{2l} S_{2l} + n_3 S_3 + n_4 S_4)/S_{2h} + E[I_{\text{intercell}}]/S_{2h} \quad , \qquad (9)$$
$$\sigma_{2h} = \frac{\sqrt{Var[I_{\text{intercell}}]}}{S_{2h}} \quad .$$

4. Interactive

$$\Pr(BER_{3} \geq BER_{3}^{*}) = \Pr(SIR_{3} \leq SIR_{3}^{*})$$
$$= \sum_{n_1=0}^{N_1} \sum_{n_{2l}=0}^{MN_2} \sum_{n_{2h}=0}^{N_2} \sum_{n_3=0}^{N_3-1} \sum_{n_4=0}^{N_4} Q\left(\frac{\delta_{3}-\mu_{3}}{\sigma_{3}}\right) \times \binom{N_1}{n_1} p_1^{n_1} (1-p_1)^{N_1-n_1} \times \binom{MN_2}{n_{2l}} p_{2l}^{n_{2l}} (1-p_{2l})^{MN_2-n_{2l}} \times$$
$$\binom{N_2}{n_{2h}} p_{2h}^{n_{2h}} (1-p_{2h})^{N_2-n_{2h}} \times \binom{N_3-1}{n_3} p_3^{n_3} (1-p_3)^{N_3-1-n_3} \times \binom{N_4}{n_4} p_4^{n_4} (1-p_4)^{N_4-n_4} \quad ,$$

where

$$\delta_3 = G_3/\gamma_3 - \eta/S_3 \quad ,$$
$$\mu_3 = n_3 + (n_1 S_1 + n_{2l} S_{2l} + n_{2h} S_{2h} + n_4 S_4)/S_3 + E[I_{\text{intercell}}]/S_3 \quad , \tag{10}$$
$$\sigma_3 = \frac{\sqrt{Var[I_{\text{intercell}}]}}{S_3} \quad .$$

5. Background

$$\Pr(BER_4 \geq BER_4^*) = \Pr(SIR_4 \leq SIR_4^*)$$
$$= \sum_{n_1=0}^{N_1} \sum_{n_{2l}=0}^{MN_2} \sum_{n_{2h}=0}^{N_2} \sum_{n_3=0}^{N_3} \sum_{n_4=0}^{N_4-1} Q\left(\frac{\delta_4 - \mu_4}{\sigma_4}\right) \times \binom{N_1}{n_1} p_1^{n_1} (1-p_1)^{N_1-n_1} \times \binom{MN_2}{n_{2l}} p_{2l}^{n_{2l}} (1-p_{2l})^{MN_2-n_{2l}} \times$$
$$\binom{N_2}{n_{2h}} p_{2h}^{n_{2h}} (1-p_{2h})^{N_2-n_{2h}} \times \binom{N_3}{n_3} p_3^{n_3} (1-p_3)^{N_3-n_3} \times \binom{N_4-1}{n_4} p_4^{n_4} (1-p_4)^{N_4-1-n_4} \quad ,$$

where

$$\delta_4 = G_4/\gamma_4 - \eta/S_4 \quad ,$$
$$\mu_4 = n_4 + (n_1 S_1 + n_{2l} S_{2l} + n_{2h} S_{2h} + n_3 S_3)/S_4 + E[I_{\text{intercell}}]/S_4 \quad , \tag{11}$$
$$\sigma_4 = \frac{\sqrt{Var[I_{\text{intercell}}]}}{S_4} \quad .$$

We solve equation (1) to get the positive solution of the powers, S_1, S_{2l}, S_{2h}, S_3, S_4, and with the above equations of outage probability of each class to obtain the final WCDMA wireless admission region.

Bibliography

[1] R. Braden, D. Clark and S. Shenker, “Integrated Services in the Internet Architecture: an Overview”, IETF RFC 1633, June 1994.

[2] R. Braden, Ed., L. Zhang, S. Berson, S. Herzog and S. Jamin, “Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification”, IETF RFC 2205, Sept. 1997.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, “An Architecture for Differentiated Service”, IETF RFC 2475, Dec. 1998.

[4] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski, “Assured Forwarding PHB Group”, IETF RFC 2597, June 1999.

[5] B. Davie, A. Charny, J.C.R. Bennet, K. Benson, J.Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu and D. Stiliadis, “An Expedited Forwarding PHB”, IETF RFC 3246, March 2002.

[6] K. Nichols, V. Jacobson and L. Zhang, “A Two-bit Differentiated Services Architecture for the Internet”, IETF RFC 2638, July 1999.

[7] V. Elek, G. Karlsson and R. Ronngren, “Admission Control Based on End-to-End Measurements”, in Proc. IEEE INFOCOM 2000, Vol. 2, pp. 623-630.

[8] G. Bianchi and N. Blefari-Melazzi, “Admission Control over Assured Forwarding PHBs: A Way to Provide Service Accuracy in a DiffServ Framework”, in Proc. GLOBECOM 2001, Vol. 4, pp. 2561-2565.

[9] F. Borgonovo, A. Capone, L. Fratta, M. Marchese and C. Petrioli, “PCP: A Bandwidth Guaranteed Transport Service for IP Networks”, in Proc. IEEE ICC 1999, Vol. 1, pp. 671-675.

[10] J. Qiu and E.W. Knightly, "Measurement-Based Admission Control with Aggregate Traffic Envelopes", *IEEE/ACM Transactions on Networking*, Vol. 9, No. 2, pp. 199-210, April 2001.

[11] C. Oottamakorn and D. Bushmitch, "A DiffServ Measurement-Based Admission Control Utilizing Effective Envelopes and Service Curves", in Proc. IEEE ICC 2001, Vol. 4, pp. 1187-1195.

[12] G. Zhang and H.T. Mouftah, "End-to-end QoS Guarantees over DiffServ Networks", in Proc. Computers and Communications 2001, pp. 302-309.

[13] D. Cavendish and M. Gerla, "Internet QoS Routing Using the Bellman-Ford Algorithm", in Proc. IFIP Conference on High Performance Networking, 1998.

[14] S.K. Agrawal and M. Krishnamoorthy, "Resource Based Service Provisioning in Differentiated Service Networks", in Proc. IEEE ICC 2001, Vol. 6, pp. 1765-1771.

[15] 3rd Generation Partnership Project (3GPP), http://www.3gpp.org

[16] 3GPP TS 23.107, "QoS Concept and Architecture", v5.7.0, Dec. 2002.

[17] 3GPP TS 23.228, "IP Multimedia Subsystem (IMS)", v6.0.1, Jan. 2003.

[18] A.M. Viterbi and A.J. Viterbi, "Erlang Capacity of a Power Controlled CDMA System", *IEEE Journal on Selected Areas in Communications*, Vol. 11, No. 6, pp. 892-900, Aug. 1993.

[19] F.Y. Li and N. Stol, "A Priority-oriented Call Admission Control Paradigm with QoS Re-negotiation for Multimedia Services in UMTS", in Proc. IEEE VTC 2001 Spring, Vol. 3, pp. 2021-2025.

[20] M. Kazmi, P. Godlewski and C. Cordier, "Admission Control Strategy and Scheduling Algorithms for Downlink Packet Transmission in WCDMA", in Proc. IEEE VTC 2000, Vol. 2, pp. 674-680.

[21] H. Holma and J. Laakso, "Uplink Admission Control and Soft Capacity with MUD in CDMA", in Proc. IEEE VTC 1999-Fall, Vol. 1, pp. 431-435.

[22] S. Akhtar, S.A. Malik and D. Zeghlache, "Prioritized Admission Control for Mixed Services in UMTS WCDMA Networks", in Proc. IEEE PIMRC 2001, Vol. 1, pp. B-133 – B-137.

[23] Z. Liu and M.E. Zarki, "SIR-Based Call Admission Control for DS-CDMA Cellular Systems", *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 4, pp. 638-644, May 1994.

[24] F. Kelly, "Notes on Effective Bandwidths", in Stochastic Networks: Theory and Applications, Oxford University Press, pp. 141-168, 1996.

[25] V.G. Kulkarni and N. Gautam, "Admission Control of Multi-Class Traffic with Service Priorities in High-Speed Networks", *Queuing Systems: Theory and Applications*, Dec. 1997, Vol. 27, pp. 79-97.

[26] A.W. Berger and W. Whitt, "Effective Bandwidths with Priorities", *Networking, IEEE/ACM Transactions on*, Vol. 6, No. 4, pp. 447-460, Aug. 1998.

[27] P. Joos and W. Verbiest, "A Statistical Bandwidth Allocation and Usage Monitoring Algorithm for ATM Networks", in Proc. IEEE ICC 1989, Vol. 1, pp. 415-422.

[28] F.C. Schoute, "Simple Decision Rules for Acceptance of Mixed Traffic Streams", Philips TDS Review, 46, 2, pp. 35-48, 1988.

[29] P. Sen, B. Maglaris, N.E. Rikli and D. Anastassiou, "Models for packet switching of variable-bit-rate video sources", *IEEE Journal on Selected Areas in Communications*, Vol. 7, No. 5, pp. 865-869, June 1989.

[30] A.I. Elwalid, D. Heyman, T.V. Lakshman, D. Mitra and A. Weiss, "Fundamental Bounds and Approximations for ATM multiplexers with Applications to Video

Teleconferencing", *IEEE Journal on Selected areas in Communications*, Vol. 13, No. 6, pp. 1004-1016, Aug. 1995.

[31] R.G. Addie, M. Zukerman and T.D. Neame, "Broadband Traffic Modeling: Simple Solutions to Hard Problems", *IEEE Communications Magazine*, Vol. 36, No. 8, pp. 88-95, Aug. 1998.

[32] S. Bodamer and J. Charzinski, "Evaluation of Effective Bandwidth Schemes for Self-Similar Traffic", in Proc. 13th ITC Specialist Seminar on IP Measurement, Modeling and Management, Vol. 21, pp. 1-10, Sept. 2000.

[33] I. Norros, "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 6, pp. 953-962, Aug. 1995.

[34] R.G. Addie, "On Weak Convergence of Long-range Dependent Traffic Processes", *Journal of Statistical Planning and Inference*, Vol. 80, pp. 155-171, 1998.

[35] Y. Jiang, C.K. Tham and C.C. Ko, "An Approximation for Waiting Time Tail Probabilities in Multiclass Systems", *IEEE Communications Letters*, Vol. 5, No. 4, pp. 175-177, Apr. 2001.

[36] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver, Jr. and C.E. Wheatley III, "On the Capacity of a Cellular CDMA System", *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 2, pp. 303-312, May 1991.

[37] R. Vannithamby and E.S. Sousa, "Performance of Multi-rate Data Traffic Using Variable Spreading Gain in the Reverse Link under Wideband CDMA", in Proc. IEEE VTC 2000 Spring, Vol. 2, pp. 1155-1159.

[38] T.C. Wong, J.W. Mark, K.C. Chua, J. Yao and Y.H. Chew, "Performance Analysis of Multiclass Services in the Uplink of Wideband CDMA", in Proc. IEEE ICCS 2002, Vol. 2, pp. 692-696.

[39] T.C. Wong, J.W. Mark, K.C. Chua and B. Kannan, "Performance Analysis of Variable Bit Rate Multiclass Services in the Uplink of Wideband CDMA", in Proc. IEEE ICC 2003, Vol. 1, pp. 363-367.

[40] T.C. Wong, J.W. Mark and K.C. Chua, "Performance Evaluation of Video Services in a multi-rate DS-CDMA System", in Proc. IEEE PIMRC 2003, Vol. 2, pp. 1490-1495.

[41] 3GPP TS 25.427, "UTRAN I_{ub}/I_{ur} Interface User Plane Protocol for DCH Data Streams", v5.1.0, Dec. 2002.

[42] 3GPP TS 25.301, "Radio Interface Protocol Architecture", v5.2.0, Sept. 2002.

[43] 3GPP TR 25.933, "IP Transport in UTRAN", v5.2.0, Sept. 2002.

[44] 3GPP TS 23.060, "General Packet Radio Service (GPRS) Service Description", v5.4.0, Dec. 2002.

[45] 3GPP TS 23.207, "End-to-End Quality of Service (QoS) Concept and Architecture", v5.6.0, Dec. 2002.

[46] C. Nie, "Packet Level Quality of Service of Multiclass Traffic in WCDMA Mobile Networks", Master Dissertation, National University of Singapore, 2003.

[47] W. Whitt, "Tail Probabilities with Statistical Multiplexing and Effective Bandwidths in Multi-Class Queues", *Telecommunication System*, Vol. 2, pp. 71-107, 1993.

[48] V.G. Subramanian and R. Srikant, "Tail Probabilities of Low-Priority Waiting Times and Queue Lengths in MAP/GI/1 Queues", *Queuing Systems : Theory and Applications*, Vol. 34, No. 1, pp. 215-236, 1998.

[49] G. Kesidis, J. Walrand and C.S. Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources", *IEEE/ACM Transaction on Networking,* Vol. 1, No. 4, pp. 424-428, 1993.

[50] J. Yao, Y.H. Chew and T.C. Wong, "Forward Link Capacity in Multi-service Cellular CDMA Systems with Soft Handoff in the Presence of Path Loss and Lognormal Shadowing", submitted to IEEE ICC 2004.

[51] L. Xiao, T.C. Wong and Y.H. Chew, "End-to-End QoS of Integrated Multi-Class Traffic in a DiffServ Network", to appear in Proc. IEEE ICCS 2004.

[52] L. Xiao, T.C. Wong and Y.H. Chew, "End-to-End QoS of Multi-Class Traffic in WCDMA and DiffServ Network", to appear in Proc. IEEE PIMRC 2004.